An Embedded Marked Point Process Framework for Three-Level Object Population Analysis

Csaba Benedek, *Member, IEEE*

***Abstract*—In this paper 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. Different from earlier, highly domain specific attempts on MPP generalization, the proposed model is defined at an abstract level, providing clear interfaces for applications in various domains. We also introduce a global optimization process for the multi-layer framework for finding optimal entity configurations, 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 and mobile laser scanning (Lidar) data and optical circuit inspection. A new benchmark database is published for the three test cases, and the model's performance is quantitatively evaluated.**

***Index Terms*—Marked point process, object population analysis, scene parsing**

I. Introduction

Object based interpretation of digital images is a crucial step in several vision applications, among others in remotely sensed data analysis, optical inspection systems, or video surveillance. Since imaging equipments are quickly improving regarding both macro and micro scale data acquisition technologies, we can witness a significant improvement of the available image resolution in many fields. Nowadays we can perceive multiple effects on different scales of a single image, thus there is a need for recognition algorithms that can perform hierarchical interpretation of the image contents [1], [2].

A widely adopted initial step towards understanding an image is to perform full-scene labeling also known as scene parsing, where we label every pixel in the image with the category of the object it belongs to [3]. Markov Random Fields (MRFs) [4] are frequently used for such tasks since the early eighties, since they are able to simultaneously embed a data model, reflecting the knowledge on the image, and prior constraints, such as the spatial smoothness of the solution through a graph based image representation. Later approaches overcome some limitations of MRFs, by allowing non-Markovian prior fields [5], or directly modeling the data-driven posterior distributions of the semantic classes, as shown in Conditional Random Fields (CRF) [6]. Recent solutions exploit deep neural networks [3] for supervised semantic segmentation. As detailed in [7], these various models realize a global scene representation based on local specifications and interactions. Although they can incorporate contextual properties in a flexible way, they prove much more limited in modeling geometric information. For example, they do not allow setting constrains on the shape of the segmentation regions without leading to prohibitive complexity, and are not well suited for the representation of macro-textures.

Marked Point Processes (MPP) [7], [8], [9] offer an efficient extension of MRFs, as they work with objects as variables instead of pixels, considering that the number of variables (i.e. objects) is also unknown. MPPs embed prior constraints and data models within the same density, therefore similarly to MRFs, algorithms for model optimization [10], [11], [12], [13] and parameter estimation [14], [15], [16] are available. Nevertheless, many available solutions are limited to specific energy functions [11], [12], [13] or use restrictive statistical assumptions such as requiring the features to be independent, having Gaussian distribution [16]. Recent MPP applications range from 2D [17] and 3D object extraction [18] in various environments, to 1D signal modeling [19] or target tracking [20], [21]. In particular, MPPs have previously been used for various population counting problems, dealing with a large number of objects which have low variation in shape, such as buildings [22], [23], trees [24], [25], birds [10], or boats [14] from remotely sensed data; road manhole and sewer well covers from Mobile Laser Scanning (MLS) measurements [26]; facial wrinkles from medical [27] and cell nuclei from biological images [28], or people in video surveillance scenarios [18], [29]. Experiments reported advantages of MPPs in population counting [7], [13] versus alternative techniques, such as Hough transform and mathematical morphology based methods [30], or the standard non-maximal suppression [13] possibly combined with window-based object detectors [31], which show limitations in cases of dense populations with several adjacent objects. MPP models can handle such phenomena more efficiently, by jointly describing individual objects with various data terms, and using information from entity interactions by prior geometric constraints [7].

As a limitation, however, classical MPP-based image anal-

Manuscript received September 13, 2016; revised February 22, 2017; accepted June 6, 2017, date of publication June, 2017; date of current version June 8, 2017. This work was supported in part by the János Bolyai Research Scholarship of the Hungarian Academy of Sciences, and in part by the Hungarian National Research, Development and Innovation Fund under Grant NKFIA K-120233. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Jianfei Cai.

C. Benedek is with the Machine Perception Research Laboratory, Institute for Computer Science and Control, Hungarian Academy of Sciences (MTA SZTAKI), H-1111 Kende u. 13-17 Budapest, Hungary and with the Faculty of Information Technology and Bionics, Péter Pázmány Catholic University, H-1083, Práter utca 50/A, Budapest, Hungary. E-mail: benedek.csaba@sztaki.mta.hu

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

ysis models [10], [13] focus purely on the object level of the scene, and they are not well suited for hierarchical pattern recognition problems in a straightforward way. Simple prior interaction constraints such as non-overlapping or parallel alignment are often utilized to refine the accuracy of the detection, but they only allow for a very limited exploitation of high level structural information from the global scene.

The Multi-MPP framework proposed by [32] offers extensions of MPP models in two senses. *First*, to simultaneously detect entities with varying shapes, it jointly samples different types of geometric objects. *Second*, local texture representations of different image regions are obtained by a statistical type and alignment analysis of nearby entities. Although this approach fits well with bottom-up exploration tasks of unknown image content, it is not straightforward how to efficiently segment the object population in such a framework, 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 artifacts included within the CEs [33], [34] in Automatic Optical Inspection (AOI) images, building roofs and chimneys in aerial or satellite photos, ships and containers in radar images [35], etc.

Up to now, only highly task specific attempts have been conducted to model the object encapsulation [33], [35] or the Bayesian object group management [36] issues within the MPP framework. Although these studies gave examples for how classical MPP schemes can be extended to solve definite issues of concrete applications, the proposed models have been investigated and evaluated purely in their original fields of application, only providing a few notes about possible generalization for different domains. Practical experiences show however, that for such complex, application dependent models, the adaptation for 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. For this reason, this paper follows a reverse path by collecting similar tasks appearing in different application areas, and addressing them by a joint methodological approach. We provide therefore a formal problem statement and introduce a novel three-level MPP framework which allows us to handle a wide family of applications. The structural elements and the energy optimization algorithm of the complex model are defined and implemented at an abstract level, while we keep focus on establishing very simple interfaces for different applications, providing efficient options for domain adaption for end-users. The proposed methodology has two key properties:

1) 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 the child objects, or the color/texture of the parent object may influence the appearance characteristics of the child entity.

2) To avoid the limitations of using only pairwise object interactions, we propose a multilevel 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.

A preliminary stage of the proposed method has been introduced in [37], [38]. This paper presents a more elaborated model with various new feature based and prior energy terms and application scenarios. We also publish a novel public benchmark for quantitative evaluation of our framework.

II. Introduction to Marked Point Processes

Similarly to Markov Random Fields (MRF) or Conditional Random Fields (CRF), Marked Point Process (MPP) methods use a graph-based representation for semantic content modeling. However, unlike in MRFs or CRFs, the graph nodes in MPPs are associated with geometric objects instead of low level pixels or 3D point cloud elements. This way an MPP model enables the characterization of whole populations instead of individual objects, by 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.

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

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

However, it is often not enough to model our objects as point-wise entities. For example, in high resolution aerial photos, building shapes can often be efficiently approximated by rectangles. To include object geometry in the model, we can assign markers to the points, for example a rectangle u can be defined by its center point $o \in S$, its orientation θ and the lengths of its perpendicular sides e_L and e_l. In this case the marker is a 3D parameter vector (θ, e_L, e_l). By 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 $\omega \in \Omega$, is a population of an unknown number of marked objects, where Ω is the population space. We also define a $\sim$ neighborhood relation between the objects of a given ω configuration: objects $u, v \in \omega$ are in a neighborhood relation $u \sim v$ iff the distance between the object centers is lower than a predefined threshold, yielding the set of $\mathcal{N}_u(\omega)$ *proximity neighborhoods* in ω.

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

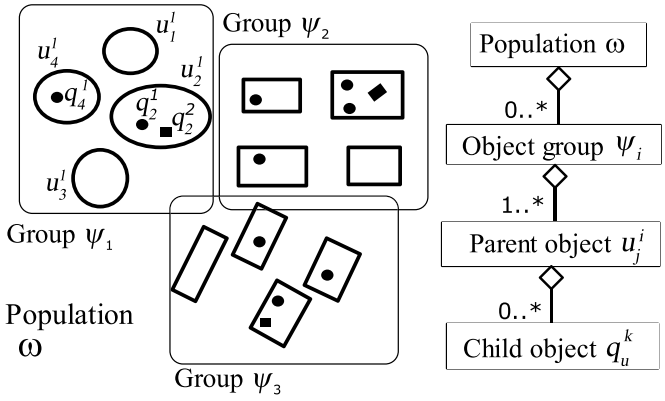


Fig. 1. Structure elements of the EMPP model. Left: a sample population with three object groups, and various object shapes both at parent and child layers. Right: The multi layer structure of the model featuring the encapsulation relation.

data-dependent Gibbs distribution over the population space:

$$P_{\mathcal{F}}(\omega) = P(\omega|\mathcal{F}) = \frac{1}{Z} \cdot \exp\Big(-\Phi(\omega)\Big) \tag{2}$$

with a Z normalizing constant: $Z = \sum_{\omega\in\Omega} \exp\left(-\Phi(\omega)\right)$. Here $\Phi(\omega)$ is called the configuration energy. Following an energy decomposition approach - also used by MRFs - we obtain $\Phi(\omega)$ as the sum of simple components, which can be calculated by considering small subconfigurations only. To obtain the optimal configuration one should minimize $\Phi(\omega)$, which can be performed with various iterative algorithms perturbating the population with preliminary defined kernels following different sampling processes [10], [16].

III. Proposed three-level EMPP framework

To model the hierarchical scene content, the proposed Embedded Marked Point Process (EMPP) framework has a multilayer structure, as shown in Fig. 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).

The input of the EMPP method is an image over a pixel lattice S, and $s \in S$ denotes a single pixel. We start with the (super) object layer, which plays a central role in the model. Let u be an object candidate of in scene, whose imaged shape is represented by a planar figure from a previously fixed shape library. In this paper ellipses ($\bigcirc$), rectangles ($\square$) and isosceles triangles ($\triangle$) are used. The shape of u is indicated by a *shape type* attribute $\mathrm{tp}(u) \in \{\bigcirc, \square, \triangle\}$. For each object, we define the coordinates of a reference point $o = [o_x, o_y]$, the global orientation $\theta \in [-90°, +90°]$, and the geometry is described by a $\mathcal{K}_{\mathrm{tp}(u)}$ shape dependent parameter set, which contains the major and minor axes for ellipses, the perpendicular side lengths for rectangles, and a side-height pair for triangles. Let us denote by $\mathcal{H}_{\mathrm{tp}(u)} = S \times [-90°, +90°] \times \mathcal{K}_{\mathrm{tp}(u)}$ the complete parameter space of an u object with type $\mathrm{tp}(u)$. The unified object space can be obtained as $\mathcal{H} = \cup_{\mathrm{tp}} \mathcal{H}_{\mathrm{tp}}$.

As a next step we formulate the superobject–subobject relation. Each parent object u may contain a set of child objects $Q_u = \{q_u^1 \dots q_u^{m(u)}\}$ where $m(u) \leq m_{\max}$, and each child is a sample from the previously defined geometric figure library $q_u^i \in \mathcal{H}_{\mathrm{tp}(q_u^i)}$. $Q_u = \emptyset$ means that u has no child. Let us denote by $\mathcal{H}_Q$ the children vector's parameter space.

For the second level of the proposed object hierarchy, we introduce the object grouping process. According to our earlier definition, a given population, denoted by ω, is a set of k *object groups* or (also referred later as *configuration segments*), $\omega = \{\psi_1, \dots, \psi_k\}$, where each group ψ_i ($i = 1 \dots k$) is a configuration of n_i objects:

$$\psi_i = \{u_1^i, \dots, u_{n_i}^i\} \in (\mathcal{H} \times \mathcal{H}_Q)^{n_i}. \tag{3}$$

Here we prescribe that $\psi_i \cap \psi_j = \emptyset$ for $i \neq j$, while the k set number and $n_1, \dots, n_k$ set cardinality values may be arbitrary (and initially unknown) integers. We denote with $u \prec \omega$ in the case when 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, constructed as:

$$\Omega = \cup_{k=0}^{\infty} \left\{ \{\psi_1, \dots, \psi_k\} \in \left[\cup_{n=1}^{\infty} \Psi_n\right]^k \right\} \tag{4}$$
$$\text{where } \Psi_n = \left\{ \{u_1, \dots, u_n\} \in (\mathcal{H} \times \mathcal{H}_Q)^n \right\}.$$

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.

IV. EMPP energy model

The EMPP framework follows an inverse modeling approach, so that an energy function $\Phi(\omega)$ is defined, which can evaluate each $\omega \in \Omega$ configuration based on the observed data and prior knowledge. Therefore, the energy can be decomposed into a unary term (Y) and an interaction term (I):

$$\Phi(\omega) = \Phi_Y(\omega) + \Phi_I(\omega), \tag{5}$$

and the optimal $\widehat{\omega}$ configuration is obtained by minimizing $\Phi(\omega)$:

$$\widehat{\omega} = \underset{\omega\in\Omega}{\operatorname{argmin}}\, \Phi(\omega). \tag{6}$$

A. Unary object appearance terms

Each object u is associated with a *unary* energy term $\varphi_Y(u)$, which characterizes u depending on the local image data, independent of other objects of the population. The unary term $\varphi_Y(u)$ is decomposed into a parent term $\varphi_Y^p(u)$ and for each child object q_u a child term $\varphi_Y^c(u, q_u)$. As indicated by the notation, the child term may depend on both the local image data and the geometry of the parent object (e.g. an intensity histogram within the parent region).

At the *parent level*, we first define different $f_i(u) : \mathcal{H} \to \mathbb{R}$ features ($f_1 \dots f_k$, $\forall i$ $f_i(u) \in [0, 1]$) which evaluate an object hypothesis for u in the image, so that 'high' $f(u)$ values correspond to effective object candidates. In the *second step*,

we construct $\phi_f(u)$ *data driven* energy subterms for each feature f, by attempting to satisfy $\phi_f(u) < 0$ for real objects and $\phi_f(u) > 0$ for false candidates. For this purpose, we project the feature domain to $[-1, 1]$ with a monotonously decreasing nonlinear $\mathcal{M}(f, d_0^f)$ function [22], whose zero value is equal to parameter d_0^f:

$$\begin{aligned}\phi_f(u) =& \mathcal{M}(f(u), d_0^f) = \\ =& \begin{cases} \left(1 - \frac{f(u)}{d_0^f}\right), & \text{if } \ f(u) < d_0^f \\ \exp\left(-\frac{f(u)-d_0^f}{0.1}\right) - 1, & \text{if } \ f(u) \geq d_0^f. \end{cases}\end{aligned} \tag{7}$$

In other words, d_0^f is the object acceptance threshold for feature f.

Usually a single image feature cannot reliably validate a hypothesis of presence or absence of a given object. We therefore established a general feature integration strategy, where we can combine various descriptors on a case-by-case basis with regard each application. The feature selection-integration process is based on the investigation of the observed feature histograms calculated for manually annotated true training objects. For features which are characteristic for the whole population (e.g. a single peak or plateau exists in the histogram), the d_0^f threshold is selected as the minimal f feature value observed among the training samples (using a tolerance factor for considering outliers). While this strategy ensures that almost all real objects that are consistent with the training set are marked as *attractive* by the $\phi_f(u)$ subterm, it may also cause a high false positive detection rate. The false hits are eliminated by simultaneously considering multiple feature constraints for acceptable objects, and by joining the corresponding feature energy subterms by the $\max$ operator, which is equivalent to the logical AND operation in the negative log-likehood domain (real objects should be attractive according to all prescribed feature constraints).

On the other hand, some useful features may only be characteristic for a segment of the population. For example certain buildings with red roofs, or yellow cabs in the traffic flow can be easily recognized through color filtering in an illuminant invariant color representation (such as in the HSV or CIE L*u*v* color spaces), but this filter will eliminate all non-red roofs, or non-yellow cars. In this case, the feature histogram derived from all training objects has multiple modes, where the first mode strongly overlaps with the background domain (e.g. gray cars cannot be distinguished from the road based on color). Therefore, we choose here a subsequent mode's lower boundary as the acceptance threshold of the selected f feature, meanwhile we consider that an object *prototype* energy function containing the ϕ_f subterm will label only a part of the possible objects as *attractive*. Nevertheless, several different object prototypes can be detected simultaneously in a given image, if the prototype-energies are joined with the $\min$ (logical OR) operator. Concrete examples for the data term construction process are provided in Sec. VI.

The construction of the *child's unary term* $\varphi_Y^c(u, q_u)$ is based on similar principles: it is obtained using different features mapped by the $\mathcal{M}$ function. The unary term of u is the sum of the parent level terms and the child level terms:

$$\varphi_Y(u) = \varphi_Y^p(u) + \sum_{q_u \in Q_u} \varphi_Y^c(u, q_u). \tag{8}$$

The data term of the whole configuration is obtained as the sum of the individual object energies:

$$\Phi_Y(\omega) = \sum_{u \prec \omega} \varphi_Y(u). \tag{9}$$

B. Interaction terms

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

$$\Phi_p(\omega) = \underbrace{\sum_{u \sim v} I(u, v)}_{\text{parent-parent interaction}} + \underbrace{\sum_{u \prec \omega} J(u, Q_u)}_{\text{parent-child interaction}} + \underbrace{\sum_{u, \psi} A(u, \psi)}_{\text{parent-group interaction}} \tag{11}$$

First, the $I(u, v)$ terms provide classical pairwise interaction constraints, e.g. they can penalize overlapping objects within the ω configuration:

$$I(u, v) = \frac{\text{Area}\{R_u \cap R_v\}}{\text{Area}\{R_u \cup R_v\}}, \tag{12}$$

where $R_u \subset S$ denotes the pixels covered by the geometric figure of u.

Second, the $J(u, Q_u)$ 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 the same shape type, similar color, size, orientation etc.

Third, with the $A(u, \psi)$ energies, one can define various constraints between the object group level and the (parent) object level of the scene. 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 neighbor within ψ w.r.t. relation $\sim$. Thus we derive a modified distance:

$$\hat{d}_\psi(u) = \begin{cases} 1 & \text{if } \nexists v \in \psi \backslash \{u\} : u \sim v \\ d_\psi(u) & \text{otherwise} \end{cases} \tag{13}$$

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

$$A(u, \psi) = c \text{ iff } \psi = \{u\}, \tag{14}$$

with a small $0 < c$ constant (used $c = 0.05$).

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:

$$A(u, \psi) = \begin{cases} \hat{d}_\psi(u) & \text{if } u \in \psi \\ 1 - \hat{d}_\psi(u) & \text{if } u \notin \psi. \end{cases} \tag{15}$$

Algorithm 1: Optimization of the configuration

Steps of the algorithm

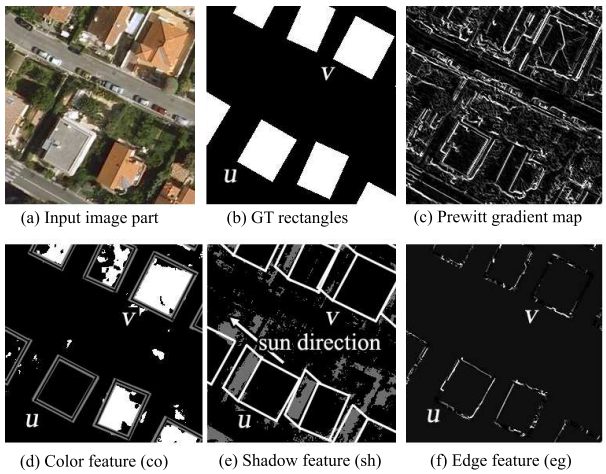
1) Initialization: start with an empty population $\omega = \emptyset$, set b_0 birth rate and $B(.)$ birth maps of the Bottom-Up Stochastic Entity Proposal (BUSEP) process, initialize the inverse temperature parameter $\beta = \beta_0$ and the discretization step $\delta = \delta_0$.
2) Main program: alternate the following three steps:
   - *Birth step*: Visit all pixels on the image lattice S sequentially. At each pixel s, with probability $\delta b_0 \cdot B(s)$, generate a new object u with center s and random geometric parameters according to the BUSEP. 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 $\varphi_Y(u) + A(u,\psi)\big|_{u\in\psi}$. For each object u taken in this order, compute $\Delta\Phi_\omega(u) = \Phi_{\mathcal{D}}(\omega/\{u\}) - \Phi_{\mathcal{D}}(\omega)$, derive the *death rate* $p_\omega^d(u)$ as
   $$p_\omega^d(u) = \Gamma(\Delta\Phi_\omega(u)) = \frac{\delta \exp(-\beta \cdot \Delta\Phi_\omega(u))}{1 + \delta \exp(-\beta \cdot \Delta\Phi_\omega(u))}, \quad (10)$$
   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, sequentially. For each object u of segment ψ we propose an alternative object u', so that the shape type of u', $\mathrm{tp}(u')$, may be different from $\mathrm{tp}(u)$, 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 $u' \in \psi'$:
   $$\Delta\varphi(\omega, u, u') = \varphi_Y(u') - \varphi_Y(u) + \sum_{v \prec \omega \setminus \{u\}} \big[I(u', v) - I(u, v)\big] + A(u', \psi') - A(u, \psi)$$
   The *object exchange rate* is calculated using the $\Gamma(.)$ function defined by (10):
   $$p_\omega^{\mathrm{e}}(u, u') = \Gamma\Big(\Delta\varphi(\omega, u, u')\Big)$$
   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.

Fig. 2. Pseudo code of Multilevel Multiple Birth and Death algorithm

V. Optimization

MPP energy functions are optimized in the literature either with stochastic iterative algorithms such as the Reversible Jump Markov Chain Monte Carlo (RJMCMC) sampler [16] and the Multiple Birth and Death Dynamic technique (MBD) [10], or with deterministic methods including the Multiple Birth and Cut algorithm (MBC) [11] and the very recent Local Submodular Approximation (LSA) [13]. The mentioned deterministic methods can provide a high quality solution with very efficient computational costs, however they are restricted to specific energy functions (singleton and doubleton terms only), and they cannot be adopted to the proposed complex three-level EMPP model in a straightforward way.

In most RJMCMC based solutions, each iteration of the relaxation consists in perturbing one or a couple of objects with various kernels such as birth, death, translation, rotation or dilation. Experiments show that the rejection rate, especially for the birth step, may induce a heavy computation time. Besides, one should decrease the temperature slowly, because at low temperatures, it is difficult to add objects to the population. On the other hand, MBD [10] evolves the population of objects by alternating purely stochastic object generation (*birth*) and removal (*death*) steps, in a Simulated Annealing (SA) framework. Each birth step of MBD consists of adding several random objects to the current configuration, and there is no rejection during the birth step, therefore high energy objects can still be added independently of the temperature parameter. Due to these properties, in several tasks a notable gain has been reported in optimization speed versus RJMCMC [10]. Note that the speed of RJMCMC can be increased with parallel implementation on GPU [12], but this solution needs partitioning the nodes into independent groups, which is not



(a) Input image part (b) GT rectangles (c) Prewitt gradient map

(d) Color feature (co) (e) Shadow feature (sh) (f) Edge feature (eg)

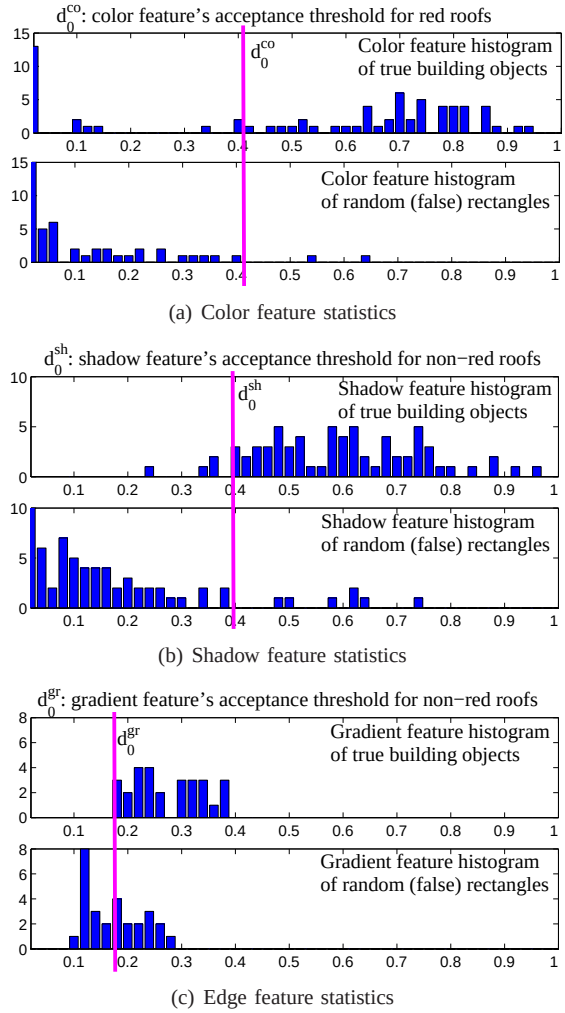
Fig. 3. Building analysis - data term features based on [22], u: efficient edge and shadow maps, weak color information. v: detection via color map

possible for EMPP due to the $A(u, \psi)$ object–group energy component.

For optimizing the energy function of Eq. (5), we have chosen the extension of the Multiple Birth and Death (MBD) [10] algorithm, as an efficient trade-off between performance and processing speed. Since the iterative MBD [10] deals with single layer MPP models, the main task here is 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 redirect some objects to neighboring object groups based on data dependent 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. The speed of the algorithm was significantly increased by the Bottom-Up Stochastic Entity Proposal (BUSEP) process [33], which assigns to the different image pixels (1) pseudo probability values of a pixel being an object reference point (e.g. center of an ellipse) (2) narrow distributions for object parameters expected in the given pixels. This way the entity proposal maintains the reversibility of the iterative evolution process of the object population [39], instead of implementing a greedy algorithm. On the other hand, this bottom-up process can efficiently guide the object exploration step towards efficient candidates. Using BUSEP we obtained the final result for each application within 30 seconds in average as detailed in Sec. VIII-C. The pseudo code of the new Multilevel Multiple Birth-Death-Maintenance (MMBDM) algorithm is shown in Fig. 2. We set the *relaxation* parameters based on [10] and used $\delta_0 = 10000$, $\beta_0 = 20$ and geometric cooling factors $1/0.96$.

VI. Applications

In this section, we introduce three different applications of the proposed EMPP model. Implementing the interfaces of the EMPP framework consists of specifying the following elements for each application:



(a) Color feature statistics

(b) Shadow feature statistics

(c) Edge feature statistics

Fig. 4. Histograms of color, shadow and edge features for true and false training objects in the built-in area analysis task

1) *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* of the same parent.
2) *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 (Sec. IV-A).
3) *Parent-parent interactions:* defining the $I(u, v)$ interaction terms between (spatially) neighboring parent objects (Sec. IV-B).
4) *Parent-child interactions:* defining the $J(u, Q_u)$ interaction constraints between the corresponding parent and children objects (Sec. IV-B).
5) *Parent-group interactions:* defining the grouping constraints through the definition of the $\hat{d}_\psi(u)$ object-segment distance (Sec. IV-B).

We would like to emphasize here that all further model elements and algorithmic steps introduced in Sections III-V are independent of any specific application.

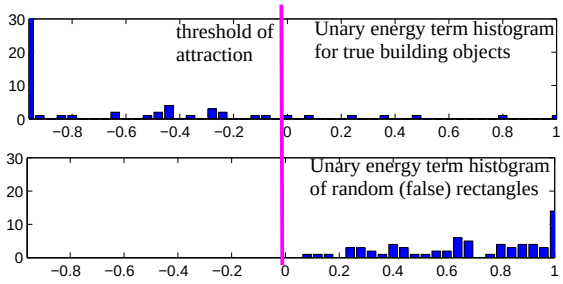


Fig. 5. Histogram of the constructed unary energy terms for true and false training objects in the built-in area analysis task

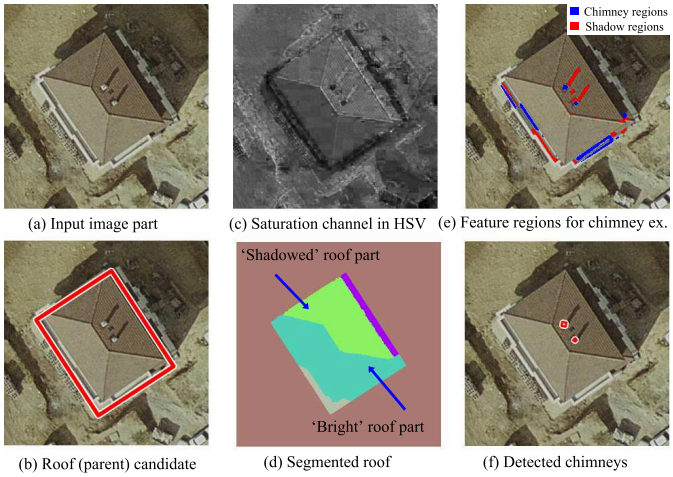


Fig. 6. Building analysis - Features for chimney extraction

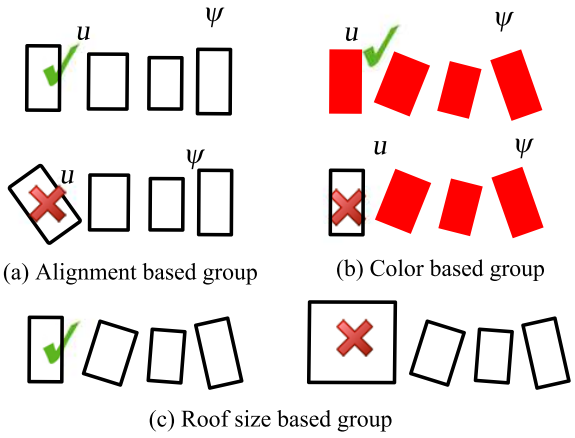


Fig. 7. Prior energies for building grouping (a)-(c) Favored ($\surd$) and penalized ($\times$) sub-configurations within a building group

A. Built-in area analysis in aerial and satellite images

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 [22], however, as pointed out in [40] 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, with a sample displayed in Fig. 9. The task specific issues are detailed in the following:

1) Model elements: 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. 9(a).

2) Parent unary terms: the $\varphi_Y^p(u)$ energy function integrates feature information about roof color, roof edge and shadow [22], as demonstrated in Fig. 3. Following the unary term construction strategy introduced in Sec. IV-A, we investigated the individual feature histograms collected from true and false training objects for feature selection and parameter estimation (see Fig. 4). *Red* roofs can be detected in color images using the hue components of the corresponding pixel values (Fig. 3(d)). The color term favors objects which contain a majority of roof colored pixels inside the rectangle of u and background pixels around u; the features are the ratios of the areas of roof-classified pixels in the internal and external boundary regions of the candidate rectangle, respectively. As shown in Fig. 4(a), the color feature histogram is multimodal, but the upper region (red roofs) can be well separated from the background using an appropriately chosen acceptance d_0^{co} threshold value (we set this threshold with the aim of minimizing the false positives).

For non-red roofs we can rely on the shadow and gradient maps [22]. As demonstrated in Fig. 3(e) the *shadowness* feature is based on a preliminary cast shadow mask, by exploiting that cast shadows are located next to the R_u object rectangles, by checking for the presence of shadows in a parallelogram T_u^{sh} defined by R_u and the estimated Sun direction vector. The *shadowness* feature is calculated as the minimum of the filling ratio of the shadowed pixels in T_u^{sh}, and the filling ratio of the non-shadowed pixels in R_u. Fig. 4(b) displays the *shadowness* histograms of true and false object candidates: the objects' domain can be well described by a lower threshold d_0^{sh}, expecting some outlier buildings, where the shadow mask could not be obtained due to background texture (such as v of Fig. 3(e)).

The *edge* descriptor exploits the information that below the edges of a relevant rectangle candidate (R_u), we expect pixels (s) with large intensity gradient vectors (∇g_s) directed towards

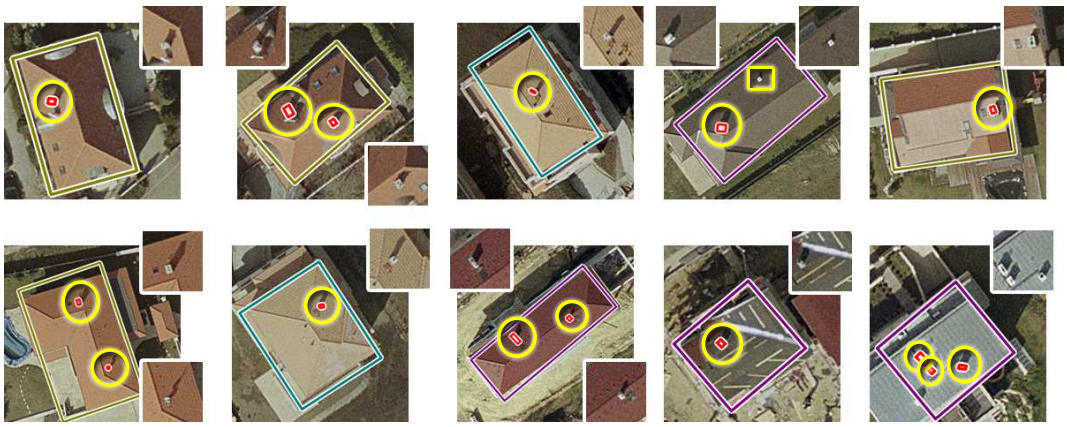


Fig. 8. Building analysis - sample results for chimney detection. True hits are marked by yellow circles, a false negative is highlighted in the fourth image of the upper row by a yellow rectangle. In the corners of the samples, the raw images of the chimney regions are displayed separately for visual verification

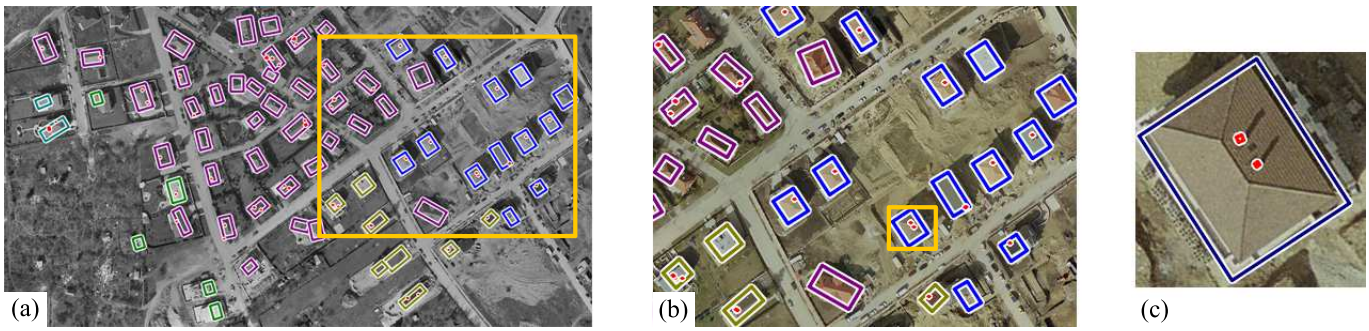


Fig. 9. Results of built-in area analysis, displayed at three different scales. Building groups are distinguished with different colors (purple: *red roofs' district*, others: orientation based groups); red markers denote the detected chimneys

to the local normal vector ($\boldsymbol{n}_s$) of the rectangle. Therefore the gradient descriptor is obtained as $\sum_{s\in\tilde{\partial}R_u} \nabla g_s \cdot \boldsymbol{n}_s$, where '$\cdot$' denotes the scalar product and $\tilde{\partial}R_u$ is the dilated edge mask of rectangle R_u (see Fig. 3(c)(f)). Edge feature histograms can be examined in Fig. 4(c).

We have empirically observed that the above three descriptors are efficient complementary features in many scenes, and we use two prototypes in the model: the first one uses the edge (eg) and shadow (sh) constraints in parallel, while the second one considers the roof color only (co). By using the ϕ_{eg}, ϕ_{sh} and ϕ_{co} primitive terms defined by Eq. (7), the joint parent level energy value is calculated as:

$$\varphi_Y^p(u) = \min\left\{ \max\left\{\phi_{\mathrm{eg}}(u), \phi_{\mathrm{sh}}(u)\right\}, \phi_{\mathrm{co}}(u)\right\}. \qquad (16)$$

Fig. 5 provides an initial validation of the above choice: histograms of the $\varphi_Y^p(u)$ values over true and false objects indicate that an efficient separation is ensured by the data term in the joint feature space.

3) Child unary terms (φ_Y^c): the feature extraction workflow for indicating chimneys (or further tall structure elements) on the roofs is demonstrated in Fig. 6. 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. 6(c)). 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 [41]) we must separately handle the cases of illuminated and self-shadowed roof segments. Taking a photometric approach [42], 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 [42], 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. 6(b)(d)), then 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. 6(e) 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. 6 and Fig. 8.

4) 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. 9c).

5) 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

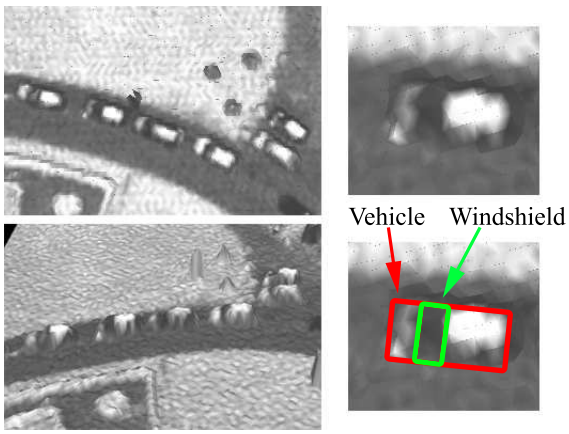


Fig. 10. Vehicles appearances in raw triangulated Lidar data (intensity based coloring was used)

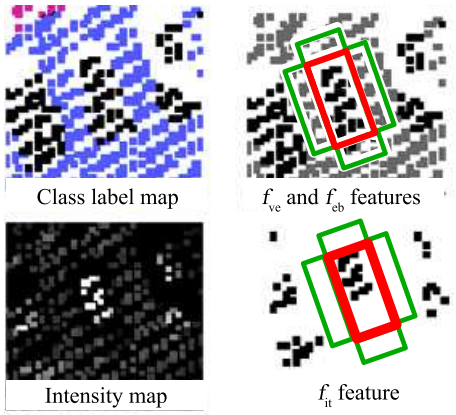


Fig. 11. Traffic monitoring application, calculation of the data model features based on [36]

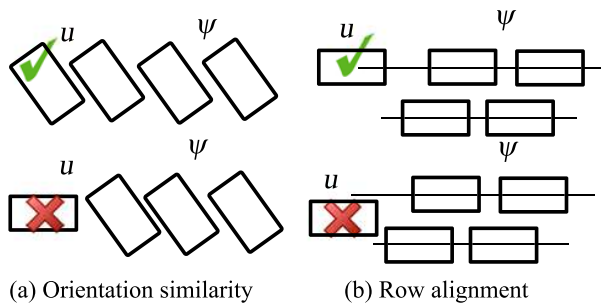
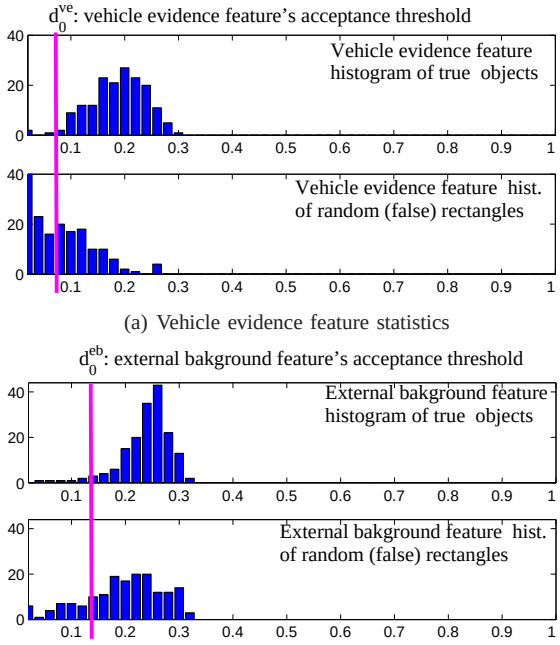


Fig. 12. Traffic monitoring application, calculation of the prior grouping features a)-b) Favored ($\surd$) and penalized ($\times$) sub-configurations within a traffic segment



(a) Vehicle evidence feature statistics

(b) External background feature statistics

Fig. 13. Histograms of *vehicle evidence* and *external background* features for true and false training objects in the traffic monitoring task

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. 9(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. 7(b)), $d_\psi(u)$ is proportional to the angle difference between u and the mean angle within ψ. Otherwise, if ψ is a color group (Fig. 7(c)), $d_\psi(u)$ 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 (Fig. 9(a),(b). Finally, for separating individual houses from larger condominiums, the roof size and the side length ratios are the discriminative features.

B. 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. 10, 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. 18), 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

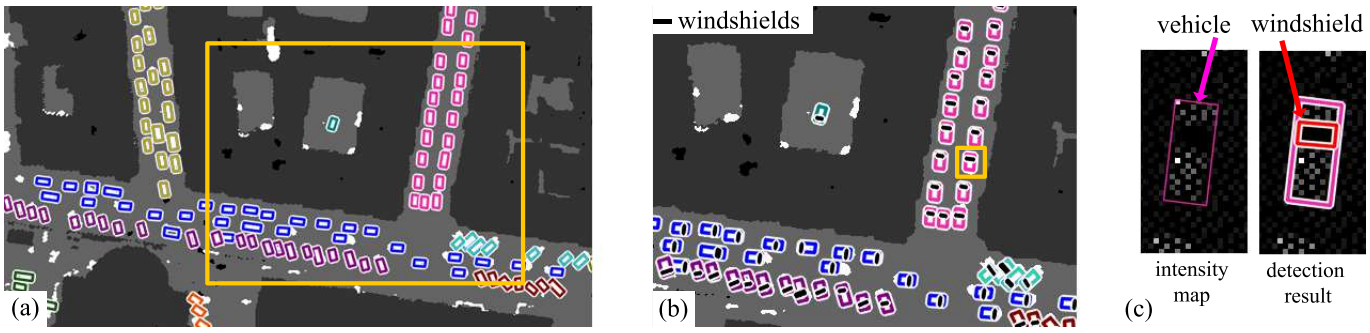


Fig. 14. Sample results on traffic analysis. Super rectangles mark the detected vehicles, different colors correspond to the different groups. In the background, gray levels refer to the input label map: white - vehicle candidates, light gray - road, dark gray - roof. a) cars and traffic segments b) selected region with the detected windshields c) intensity map of a selected car, d) detection result for c).

process.

In [36] a two-step method was introduced for Lidar based vehicle detection, which we adapt and extend 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. 14(a), while the projected intensity map of an MLS data segment is shown in Fig. 18(c).

1) 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. 14(a)).

2) Parent unary terms (φ_Y^p): similarly to [36], three different features are exploited for vehicle extraction (see Fig. 11). 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. (7), similarly to the built-in area analysis application (Sec. VI-A2). Finally the joint data energy of object u is calculated as:

$$\varphi_Y^p(u) = \max(\min(\phi_{\mathrm{it}}(u), \phi_{\mathrm{ve}}(u)), \phi_{\mathrm{eb}}(u)), \qquad (17)$$

where we admit that not necessarily all vehicles appear as bright blobs in the intensity map. For demonstrating the parameter choice, feature histograms of the vehicle evidence and external background descriptors are shown in Fig. 13.

3) 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. 10 and 14(c)), features which are characterized by coverage ratios similarly to the parent level descriptors.

4) 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. 14(c)).

5) 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. 12(a)). 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. 12(b)). A generalization of this feature for curved road segments can be found in [36].

C. 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 [43]. Here a common problem, called *scooping* may occur during manufacturing, which influences the strength of solder joints in stencil prints [33]: 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. 16(a). Automatic detection is challenging due to the locally varying contrast of AOI images [33].

1) 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.

2) Parent unary terms (φ_Y^p): In the considered PCB image data set [34] 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 [10] distance $d_B(u)$ between the pixel intensity distributions of the internal CE regions and

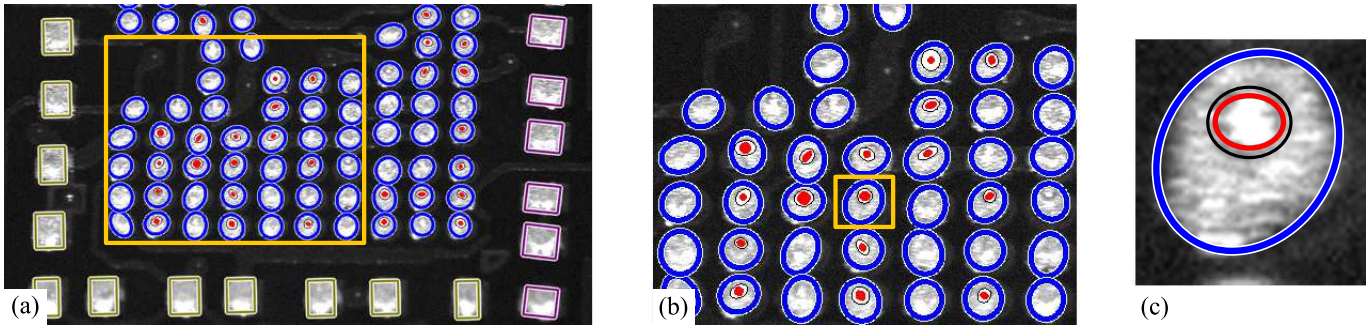


Fig. 15. Results of PCB analysis. CEs are grouped by shape and orientation, scoops are extracted within the CEs

their boundaries (see Fig. 16(a)). Then the $\varphi^p_Y(u)$ unary term is derived by $\mathcal{M}$ mapping of $d_B(u)$ (Sec. IV-A).

3) Child unary terms (φ^c_Y): Following the approach of [33] we distinguish three regions of each scoop: the central bright ellipse, the darker median ring and the bright external ring, as shown in Fig. 16(b). Experimental evidences prove (Fig. 16(c)), that for a real scoop q, the gray level histogram of the central region, $\lambda^{\mathrm{c}}_q(x)$ follows a skewed distribution, while the medium and external region histograms ($\lambda^{\mathrm{m}}_q(x)$ resp. $\lambda^{\mathrm{e}}_q(x)$) can be approximated by Gaussian densities. Let us denote by μ^{c}_q, μ^{m}_q resp. μ^{e}_q the peak locations of the smoothed $\lambda^{\mathrm{c}}_q(x)$, $\lambda^{\mathrm{m}}_q(x)$ resp. $\lambda^{\mathrm{e}}_q(x)$ functions. We prescribe three constraints for an efficient scoop candidate: (i) it exhibits high μ^{c}_q value; while intensity ratios (ii) $\mu^{\mathrm{c}}_{q_u}/\mu^{\mathrm{m}}_{q_u}$ resp. (iii) $\mu^{\mathrm{e}}_{q_u}/\mu^{\mathrm{m}}_{q_u}$ 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. (7):

$$\varphi^c_Y(u,q_u) = \max\Big(\mathcal{M}(\mu^{\mathrm{c}}_{q_u}, d^c), \mathcal{M}(\mu^{\mathrm{c}}_{q_u}/\mu^{\mathrm{m}}_{q_u}, d^{\mathrm{cm}}), \mathcal{M}(\mu^{\mathrm{e}}_{q_u}/\mu^{\mathrm{m}}_{q_u}, d^{\mathrm{em}})\Big) \quad (18)$$

4) 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.

5) 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. 17). 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. VI-B by the traffic monitoring application.

VII. Quantitative Evaluation Framework

Utilizing relevant test data and efficient quantitative evaluation metrics are key points in experimental method validation. 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 EMPP Benchmark database, which is designed for the evaluation of multilevel object population analysis techniques on high resolution images.

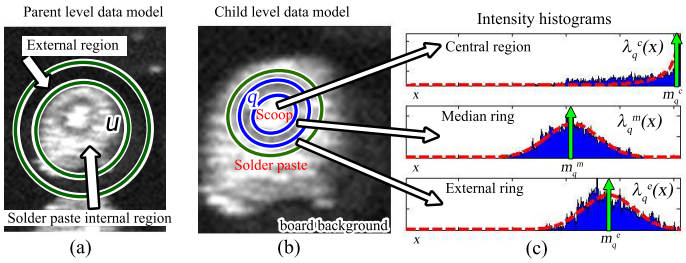


Fig. 16. Circuit inspection, calculation of the data model features based on [33]

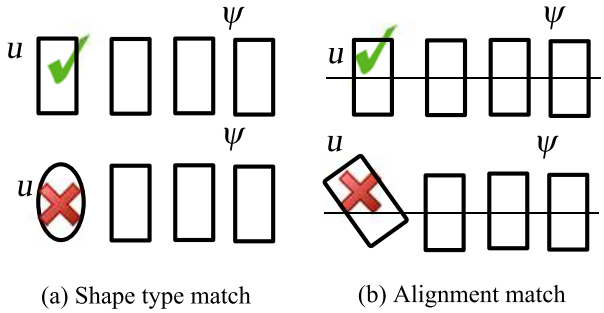


Fig. 17. Circuit inspection application, calculation of the prior grouping features (a)-(b) avored ($\surd$) and penalized ($\times$) sub-configurations within a CE group, w.r.t. the *shape type match* and *alignment match* constraints

Based on the Ground Truth (GT) data of the new benchmark, we elaborated an automatic validation methodology, which evaluates a given output configuration by comparing it to the GT, and calculates matching scores at various levels.

A. EMPP Benchmark database

The proposed EMPP Benchmark[1] is based on various (in the most part unpublished) data collections. For each scenario new Ground Truth (GT) data has been generated to enable the validation of the proposed three-layer embedded model. The serialized GT annotations encode the dependencies of objects, object groups and child objects within a population, using the same data structure and syntax for each application. (The semantic interpretation of the model elements is obviously different for each field, as introduced one-by-one in Sec.

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

TABLE I
DATASET PARAMETERS

Applicat.	Input	Resolution	Images/ scenes	Covered area	Parent objects	Child objects	Child/ parent	Group num/image
Building analysis	Rem.sens. RGB image	0.12-0.8m /pixel	4	1.0km^2	442 buildings	79* chimneys	$\{0, 1, 2, \ldots\}$	5-16
Traffic analysis (aerial/ground based)	Aerial Lidar pointcloud	8pts/m^2	6	0.3km^2	817 vehicles	817 windshields	1	7-9
	Mobile laser scan. data	up to 7000 pts/m^2	2	5700m^2	42 vehicles	42 windshields	1	3-5
PCB inspection	Grayscale AOI image	6 μm/pix	44	1232mm^2	4439 circ.elem.	664 scoops	$\{0, 1\}$	3-7

*chimneys can only be reliably analyzed in the 12cm resolution sample.

VI.) For GT annotation we have developed a program with graphical user interface, which enables us 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 (see also Table I):

1) *Building detection:* Budapest aerial image with 12cm resolution (69 buildings, 79 chimneys), Manchester satellite image (50cm res., 155 buildings) from the SZTAKI-INRIA Benchmark [22], and two Quickbird images (#2 and #11, 60cm-80cm res., 218 buildings) from the dataset by A.O. Ok [44].
2) *Traffic analysis:* the dataset contains aerial Lidar point clouds, and from a smaller region mobile laser scanning (MLS) data samples (for proof-of-concept evaluation)
   - *Aerial data:* 6 point cloud segments from Budapest, Hungary, dense urban regions, 792 vehicles (scanner: Optec ALTM Gemini 167, point denisty: 8 pts/m2) [36].
   - *MLS data:* 2 point cloud segments from Budapest, Hungary, dense urban regions, 42 vehicles (scanner: Riegl VMX-450 mobile mapping system).
3) *Optical circuit board analysis:* 44 printed circuit board images of 6μm resolution, containing 4439 CEs and 664 scooping errors [33].

B. Quantitative evaluation methodology

The quantitative evaluation of an EMPP based scene analysis algorithm should be accomplished at multiple levels. For the different layers of the model, different quality measures are defined, which can be derived fully automatically from the EMPP detection results and the GT.

In the *parent object* layer, we define both object based and pixel based accuracy rates. 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) [18], [45]. A detected object is labeled as True Positive (TP), if the HA matches 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. At the pixel level, we compare the object silhouette masks to the GT masks, and calculate the Parent Pixel level F-rate (PPF) of the match as the harmonic mean of Pixel level Precision (PPr) and Recall (PRc) [22].

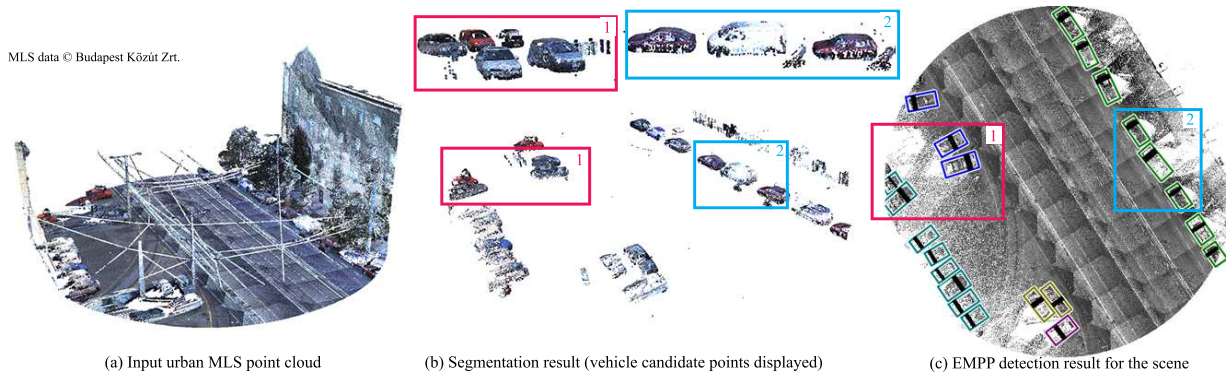
The evaluation step regarding the the *child layer* uses object level metrics similar to the parent layer. However, by calculating the Child Object level Precision (CPr), Recall (CRc) and F-rate (COF), we only accept matches between the detected and GT child objects, if their parents are also correctly 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).

VIII. EXPERIMENTS

We evaluated our method on the new EMPP Benchmark database. Qualitative sample results of the three level population detection are shown in Fig. 9, 14, 15 and 18. 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 II.

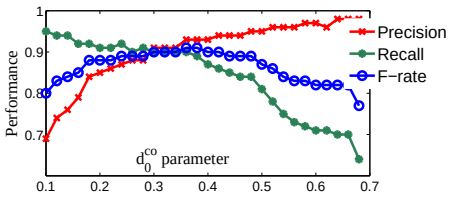
A. Performance comparison against baselines

During 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

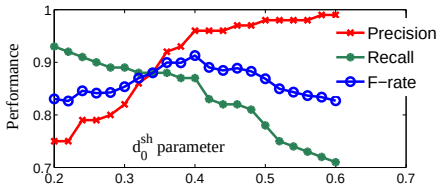


(a) Input urban MLS point cloud (b) Segmentation result (vehicle candidate points displayed) (c) EMPP detection result for the scene

Fig. 18. Processing workflow for Mobile Laser Scanning data. (a) Input scene (b) estimated vehicle regions by point cloud classification - two selected segments are highlighted from different viewpoints (c) EMPP detection results

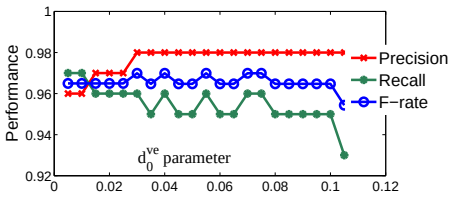


(a) Color

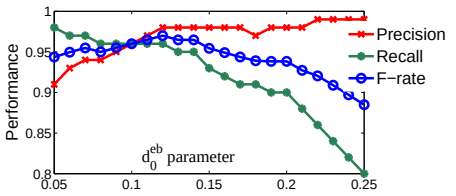


(b) Shadow

Fig. 19. Effects of the change in the data term acceptance threshold values on the object level performance for the built-in area analysis task



(a) Vehicle evidence



(b) External Background

Fig. 20. Effects of the change in the data term acceptance threshold values on the object level performance for the aerial traffic monitoring task

spatial neighbors to the same cluster iff the difference between the 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 II 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 II, 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 [32]. 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.

While the justification of using an MPP approach versus various alternative techniques for the selected application domains has already been addressed by field specific studies [22], [33], [36], we provide in Table IV a short comparison of the *parent-object level* performance of the EMPP model against various non-MPP based state of the art solutions. As references, in the building detection detection tasks we have

chosen the following three techniques: Gabor Filter based approach of [46], the Segment-Merge (SM) technique of [47], and the Orientation Selective Building Detection (OS) method proposed by [48]. For aerial vehicle detection, we compared our solution to the digital elevation map based PCA [49], h-maxima suppression (h-max) [50] and Floodfill (FF) [36]. As Table IV confirms our approach surpasses the baselines for vehicle detection task with a notable margin, while it is also competitive versus the reference techniques on building detection, overtaken only by the very recent OS with 1%.

The gain obtained by the *stochastic parent-child relationship* model of the EMPP is demonstrated in the PCB inspection application. As a baseline technique for scooping detection, we have implemented a morphology-based solution called *Morph* (introduced in [34]), 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 V shows the scooping detection performance of the deterministic *Morph* and the stochastic EMPP approach: 20% gain can be reported for EMPP at the child level.

B. Effects on data term parameter settings

As discussed in Sec. IV and VI, the most important application dependent parameter, which significantly affects the performance of the method, is the d_0^f object acceptance threshold value associated with the different features in the $\varphi_Y(u)$ unary term (and similar thresholds of the child-data terms). Fig. 4 and 13 already demonstrated the importance of appropriate d_0^f selection in discriminating real objects from false object candidates. Note that the interaction term $I(u,v)$ of Eq. (11) has a non-maxima suppression effect by removing object candidates strongly overlapping with objects having lower $\varphi_Y(u)$ unary energies, therefore several suboptimal attractive objects will not appear as false detections. For investigating the performance dependence of the complete method on the data term threshold parameter, we plot in Fig. 19 and 20 the measured *object level* precision, recall and F-rate values as a function of the d_0^f parameters corresponding to four selected feature regarding the built-in analysis and traffic monitoring tasks, respectively. We can observe that in all cases the precision and recall curves show a nearly monotonous increasing and decreasing characteristics, respectively, since we are dealing with fitness-like $f(u)$ features, where 'high' f values indicate efficient object candidates. On the other hand, the F-rate plots are gentle curves over with a single global maximum, ensuring graceful degradation in case of minor inaccuracies of the d_0^f parameter's optimization.

C. Computational time

For keeping the computational time of the iterative MMBDM optimization algorithm low, we applied an exponential temperature cooling strategy, and took the advantage of the Bottom-Up Stochastic Entity Proposal (BUSEP) process (from Sec. V), by using various application-dependent image features [22], [33], [36]. This way, the algorithm converged quickly to a sub-optimal solution, which proved to be efficient in all application domains, as demonstrated in Sec. VIII-A. 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. III 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.

D. Experiment repeatability

The iterative Multilevel Multiple Birth and Death optimization algorithm detailed in Fig. 2 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. Although our experiments supported that the outputs of the proposed framework are stable – i.e., the output configurations are largely similar for each run –, 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 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 VI. 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 VII displays 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.

IX. Conclusion

This paper 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.

TABLE II
OBJECT, GROUP AND CHILD LEVEL EVALUATION OF THE THE PROPOSED EMPP MODEL, AND COMPARISON TO A CONVENTIONAL *s*MPP APPROACH

Application	Method	Parent level analysis						Group level study		Child level study		
		Number of objects			Pixel level %			Obj mis-grouping		Detection rates %		
		TP	FP	FN	PRc	PPr	**PPF**	FG#	GR%	CRc	CPr	**COF**
Building analysis	*s*MPP	406	24	36	80	75	**78**	58	14	80	71	75
	EMPP	417	14	25	84	88	**86**	28	7			
Aerial traffic monitoring	*s*MPP	792	30	25	79	77	**78**	202	25	92	92	92
	EMPP	793	30	24	82	85	**83**	43	5			
Ground-based traffic analysis	*s*MPP	42	0	0	92	86	**89**	2	5	93	93	93
	EMPP	42	0	0	96	89	**92**	0	0			
PCB inspection	*s*MPP	4408	39	31	87	86	**87**	448	10	91	95	93
	EMPP	4415	9	24	92	97	**94**	137	3			

TABLE III
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

TABLE IV
COMPARISON OF PARENT OBJECT-LEVEL F-RATES BETWEEN VARIOUS BUILDING AND VEHICLE DETECTION TECHNIQUES

Building*	Gabor [46]	SM [47]	OS [48]	EMPP
	83%	92%	97%	96 %
Vehicle**	PCA [49]	h-max [50]	FF [36]	EMPP
	80%	83%	86%	96 %

*on the Budapest image, **complete aerial Lidar dataset

TABLE V
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-rate
Morph technique [34]	514	228	150	73%
Proposed EMPP	629	65	35	93%

TABLE VI
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 VII
DISTRIBUTION OF THE NUMBER OF FALSELY GROUPED OBJECTS (OUT OF 169 VEHICLES) IN THE 200-RUN EXPERIMENT OF TABLE VI

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

ACKNOWLEDGMENT

The author gratefully acknowledges the access to the test data: the Budapest image from András Görög, satellite images from the SZTAKI-INRIA benchmark [22] and from the dataset by Ali Özgün Ok and Andrea Manno-Kovács [44], aerial Lidar point clouds from Airbus D&S Hungary, mobile laser scanning (MLS) data from Budapest Közút Zrt. and printed circuit board images from the Department of Electronic Technology, Budapest University of Technology and Economics.

REFERENCES

[1] 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.

[2] 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.

[3] C. Farabet, C. Couprie, L. Najman, and Y. LeCun, "Learning hierarchical features for scene labeling," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 8, pp. 1915–1929, Aug 2013.

[4] X. Descombes, R. Morris, J. Zerubia, and M. Berthod, "Estimation of Markov random field prior parameters using Markov chain Monte Carlo maximum likelihood," *IEEE Trans. on Image Processing*, vol. 8, no. 7, pp. 954–963, 1999.

[5] S. Derrode and W. Pieczynski, "Signal and image segmentation using Pairwise Markov chains," *IEEE Trans. on Signal Processing*, vol. 22, no. 1, pp. 2477–2489, 2004.

[6] J. D. Wegner, R. Hansch, A. Thiele, and U. Soergel, "Building detection from one orthophoto and high-resolution insar data using conditional random fields," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 4, no. 1, pp. 83–91, March 2011.

[7] F. Chatelain, X. Descombes, F. Lafarge, C. Lantuejoul, C. Mallet, R. Minlos, M. Schmitt, M. Sigelle, R. Stoica, and E. Zhizhina, *Stochastic geometry for image analysis*, ser. Digital Signal and Image Processing, X. Descombes, Ed. Wiley-ISTE, 2011.

[8] A. Baddeley and M. Van Lieshout, "Stochastic geometry models in high-level vision," *Journal of Applied Statistics*, vol. 20, pp. 231–256, 1993.

[9] M. Van Lieshout, *Markov point processes and their applications*. London: Imperial College Press, 2000.

[10] 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.

[11] 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, 2010, pp. 35–42.

[12] Y. Verdié and F. Lafarge, "Detecting parametric objects in large scenes

by Monte Carlo sampling," *International Journal of Computer Vision*, vol. 106, no. 1, pp. 57–75, 2014.
[13] T. Pham, S. Rezatofighi, I. Reid, and T. J. Chin, "Efficient point process inference for large-scale object detection," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016, pp. 2837–2845.
[14] 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 *Proc. ISPRS Technical Commission III Symposium on Photogrammetry Computer Vision and Image Analysis (PCV)*, Paris, France, 2010.
[15] C. Meillier, F. Chatelain, O. Michel, and H. Ayasso, "Nonparametric bayesian extraction of object configurations in massive data," *IEEE Trans. on Signal Processing*, vol. 63, no. 8, pp. 1911–1924, April 2015.
[16] 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.*, ser. Lecture Notes in Computer Science, Bonn, Germany, 2009, vol. 5681, pp. 221–234.
[17] 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, 2011, pp. 3628–3631.
[18] Á. Utasi and C. Benedek, "A Bayesian approach on people localization in multi-camera systems," *IEEE Trans. on Circuits and Systems for Video Technology*, vol. 23, no. 1, pp. 105–115, 2013.
[19] 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.
[20] 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.
[21] P. Craciun, M. Ortner, and J. Zerubia, "Joint detection and tracking of moving objects using spatio-temporal marked point processes," in *IEEE Winter Conference on Applications of Computer Vision*, Jan 2015, pp. 177–184.
[22] C. Benedek, X. Descombes, and J. Zerubia, "Building development monitoring in multitemporal remotely sensed image pairs with stochastic birth-death dynamics," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 1, pp. 33–50, 2012.
[23] 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.
[24] 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 Recognition*, 2011, pp. 623–627.
[25] 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, 2012, pp. 140–145.
[26] Y. Yu, J. Li, H. Guan, C. Wang, and J. Yu, "Automated detection of road manhole and sewer well covers from mobile LiDAR point clouds," *IEEE Geoscience and Remote Sensing Letters*, vol. 11, no. 9, pp. 1549–1553, Sept 2014.
[27] S. G. Jeong, Y. Tarabalka, and J. Zerubia, "Marked point process model for facial wrinkle detection," in *IEEE International Conference on Image Processing (ICIP)*, Oct 2014, pp. 1391–1394.
[28] N. J. Gadgil, P. Salama, K. W. Dunn, and E. J. Delp, "Nuclei segmentation of fluorescence microscopy images based on midpoint analysis and marked point process," in *IEEE Southwest Symposium on Image Analysis and Interpretation (SSIAI)*, March 2016, pp. 37–40.
[29] W. Ge and R. Collins, "Marked point processes for crowd counting," in *IEEE Conference on Computer Vision and Pattern Recognition*, Miami, FL, USA, 2009, pp. 2913–2920.
[30] P. Soille, *Morphological Image Analysis: Principles and Applications*, 2nd ed. Secaucus, NJ, USA: Springer-Verlag New York, Inc., 2003.
[31] L. Zhang and R. Nevatia, "Efficient scan-window based object detection using GPGPU," in *IEEE Conference on Computer Vision and Pattern Recognition Workshops*, June 2008, pp. 1–7.
[32] F. Lafarge, G. Gimel'farb, and X. Descombes, "Geometric feature extraction by a multimarked point process," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 32, no. 9, pp. 1597 –1609, 2010.
[33] C. Benedek, O. Krammer, M. Janóczki, and L. Jakab, "Solder paste scooping detection by multi-level visual inspection of printed circuit boards," *IEEE Trans. on Industrial Electronics*, vol. 60, no. 6, 2013.
[34] 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.
[35] C. Benedek and M. Martorella, "Moving target analysis in ISAR image sequences with a multiframe marked point process model," *IEEE Trans. on Geoscience and Remote Sensing*, vol. 52, no. 4, pp. 2234–2246, 2014.
[36] A. Börcs and C. Benedek, "Extraction of vehicle groups in airborne lidar point clouds with two-level point processes," *IEEE Trans. on Geoscience and Remote Sensing*, vol. 53, no. 3, pp. 1475–1489, March 2015.
[37] C. Benedek, "A two-layer marked point process framework for multi-level object population analysis," in *International Conference on Image Analysis and Recognition (ICIAR)*, ser. Lecture Notes in Computer Science, Póvoa de Varzim, Portugal, 2013, vol. 7950, pp. 160–169.
[38] ——, "Hierarchical image content analysis with an embedded marked point process framework," in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Florence, Italy, 2014.
[39] 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.
[40] A. Kovács and T. Szirányi, "Orientation based building outline extraction in aerial images," in *XXII. ISPRS Congress*, ser. ISPRS Annals Photogram. Rem. Sens. and Spat. Inf. Sci., Melbourne, Australia, 2012, vol. I-7, pp. 141–146.
[41] 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.
[42] C. Benedek and T. Szirányi, "Bayesian foreground and shadow detection in uncertain frame rate surveillance videos," *IEEE Trans. Image Processing*, vol. 17, no. 4, pp. 608–621, 2008.
[43] 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.
[44] 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, no. 10, pp. 2140–2144, Oct 2015.
[45] H. W. Kuhn, "The Hungarian method for the assignment problem," *Naval Research Logistic Quarterly*, vol. 2, pp. 83–97, 1955.
[46] 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.
[47] 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, 2005, pp. 143–148.
[48] A. Manno-Kovács and T. Szirányi, "Orientation-selective building detection in aerial images," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 108, pp. 94 – 112, 2015.
[49] Á. Rakusz, T. Lovas, and Á. Barsi, "Lidar-based vehicle segmentation," in *XX. ISPRS Congress*, ser. ISPRS Archives Photogram. Rem. Sens. and Spat. Inf. Sci., Istanbul, Turkey, 2004, vol. XXXV-2, pp. 156–159.
[50] 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.

Csaba Benedek received the M.Sc. degree in computer sciences in 2004 from the Budapest University of Technology and Economics (BME), and the Ph.D. degree in image processing in 2008 from the Péter Pázmány Catholic University, Budapest. Between 2008 and 2009 he worked as a postdoctoral researcher with the Ariana Project Team at INRIA Sophia-Antipolis, France. He is currently a senior research fellow with the Machine Perception Research Laboratory, at the Institute for Computer Science and Control of the Hungarian Academy of Sciences (MTA SZTAKI) and an associate professor with the Péter Pázmány Catholic University. He has been the manager of various national and international research projects in the recent years. His research interests include Bayesian image and point cloud segmentation, object extraction, change detection, and GIS data analysis.