2016 23rd International Conference on Pattern Recognition (ICPR)
Cancún Center, Cancún, México, December 4-8, 2016

A Future for Learning Semantic Models of Man-Made Environments

Wolfgang Förstner
University of Bonn, Department of Geodesy and Geoinformation
D-53121 Bonn, Nussalllee 15, Email: wf@ipb.uni-bonn.de

***Abstract*—Deriving semantic 3D models of man-made environments hitherto has not reached the desired maturity which makes human interaction obsolete. Man-made environments play a central role in navigation, city planning, building management systems, disaster management or augmented reality. They are characterised by rich geometric and semantic structures. These cause conceptual problems when learning generic models or when developing automatic acquisition systems. The problems appear to be caused by (1) the incoherence of the models for signal analysis, (2) the type of interplay between discrete and continuous geometric representations, (3) the inefficiency of the interaction between crisp models, such as partonomies and taxonomies, and soft models, mostly having a probabilistic nature, and (4) the vagueness of the used notions in the envisaged application domains. The paper wants to encourage the development and learning of generative models, specifically for man-made objects, to be able to understand, reason about, and explain interpretations.**

I. Introduction

Deriving semantic 3D models of man-made environments has gained interest since the beginning of image analysis, see [1, 2] and the surveys for outdoor and indoor environments in [3, 4]. Man-made environments play a central role in navigation, city planning, building management systems, disaster management or augmented reality.

Automatic methods for semantic building reconstruction hitherto have not reached the desired maturity which makes human interaction obsolete. In spite of great success in automatically reconstructing the geometry of buildings it appears that the rich geometric and semantic structures, which characterize man-made objects, slows down progress. The paper identifies successes and difficulties in using explicit models for supporting the geometric and semantic reconstruction of buildings. We want to encourage the development and learning of generative models, specifically for man-made objects, be able to understand, reason about, and explain interpretations of man-made scenes, quite in the spirit of [5].

Based on experiences in our research group, we will discuss typical tasks which we solved using structural descriptions (image orientation, building reconstruction, and façade interpretation), and embed the used methods in the stream of concurrent solutions. We try to identify the attempts to learn the underlying models and the achievements in object recognition which on one had promise to support future methods for interpreting images of man-made scenes. However, these models – in our view – still contain conceptual problems when learning generative models or when developing automatic acquisition systems. The problems appear to be caused by (1) the incoherence of the models for signal analysis, (2) the type of interplay between discrete and continuous geometric representations, (3) the inefficiency of the interaction between crisp models, such as partonomies and taxonomies, and soft models, mostly having a probabilistic nature, and (4) the vagueness of the used notions in the envisaged application domains. A goal for future research should be to learn building models, i.e., to learn geometric, structural and semantic models which help understanding images of man-made scenes, and to further develop methods to learn these highly structured models.

We start with experiences with structural descriptions for solving tasks related to man-made objects.

II. Using Structural Descriptions

In the following we discuss three basic problems, pose determination, building reconstruction, and image interpretation in the context of man-made scenes. Pose determination is representative for the large class of parameter estimation problems based on correspondences, where – depending on the number of available image features – structural descriptions may be of advantage. Building reconstruction is a representative for the large class of problems where, besides a large number of parameters, also the structure of the solution, especially the number of parameters and possibly the constraints between the parameters, is not known from the beginning. Finally, image interpretation aims at a semantic description, thus above parameters and structure also aims at finding the class memberships of the objects and possibly the semantic relations between the objects shown in the images. The discussion of these tasks is triggered by own research and the solutions known before and achieved later and gives insight into the development during the last three decades w.r.t. the used representations and reasoning methods.

A. Relational Matching using Edges for Pose Determination

Pose estimation requires correspondences between images and a 3D model, which, when performed automatically, requires adequate matching techniques. Matching a given model with an image is based on a common representation. Thirty years ago, due to limited computer power, representations based on point or line type features dominated. Keypoints were mainly used for image-to-image matching whereas model-to-image matching mainly use image edges, even only straight

978-1-5090-4846-5/16/$31.00 ©2016 IEEE

edge segments, see e.g., [6, 7] and Fig. 1. The search for

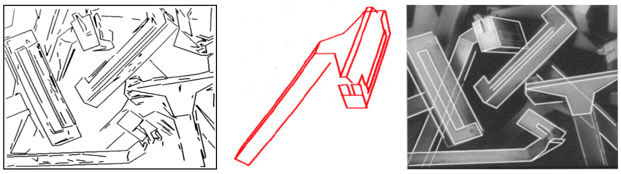

Fig. 1. Pose estimation based on straight line segments. **Left** Image edges. **Mid:** One of the models given as 3D line segments. **Right:** Image with projected model; from [7]. The matching is based on triplets of corresponding lines, which allow to directly derive the pose parameters, which then are checked for consistency with the other edges

correspondences was done incrementally, formalized as interpretation tree by [8].

Based on work on the consistent labelling problem [9, 10] and relation matching [11], we in the late nineteen eighties explored model-to-image matching for finding buildings (roofs) in aerial images [12, 13], and more general relational descriptions for pose estimation [14] or map-to-image matching based on road networks [15, 16], see Fig. 2.

Given a wire frame model of the object and line segments together with their mutual relations, such as connectivity or parallelity, the task was to derive the six parameters of the pose. Matching of the two relational descriptions using heuristic search (A*) was based on a probabilistic model of the projection. The matching costs were based on the mutual self-information $I(x;y) = -\log(P(x)/P(x|y))$.[1] The goal of the search was to maximize the sum of the mutual self-information of all matches and relations. The probabilities were learned from training data. This simplifies the evaluation of missing correspondences – often called *wild cards* in matching – by setting $I(x;y) = 0$, since the missing match has no influence. An example for detecting a road junction in an aerial image is given in Fig. 2

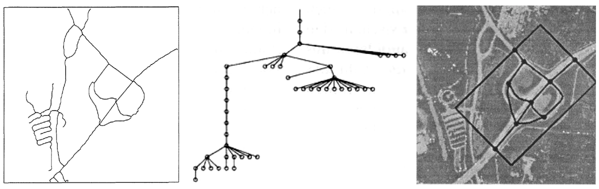

Fig. 2. Image-to-model matching. **Left:** Road map as planar graph. **Mid:** Search tree for image orientation. **Right:** Match with aerial image; from [16]. The model has 25 units (junctions, edges) (only the region around the road intersection), the image has 21 units, the search tree has 52 nodes, determining the orientation was tried six times, the software was written in POP-11, the computing time was 227 seconds on a VAX 3200

Progress in pose estimation is based on more informative features [17] or first estimating viewpoints using a regression convolutional neural network and then using key points for fine matching [18]. While both directions do not use an explicit model of the scene, exploiting a hierarchical object model for efficient detection [19], see Fig. 3 and generalizations to articulated objects are indispensable for locating persons in general pose, see e.g., [20]. Progress is triggered by a 3D recognition challenge [21].

Fig. 3. Hierarchical model for object detection, including a step for determining the orientation of the object. **Left:** Hierarchical model with three layers. **Mid:** Given image. **Right:** Bounding box, class, and projected coarse model; from [19]

B. Generic Building Models from Multiple Images using Constraint Programming

Reconstructing generic building models from images requires an adequate representation of the structure. Structure refers to the number of building parts, their relations w.r.t. neighbourhood and geometry, and to constraints between the parts, especially among parameters of the individual parts. In a first step the reconstruction only aims at a rich geometric description, and does not include an interpretation. This may be fruitful in a later step, see [22].

Early work [6] fixed the structure and only allowed variations for parameters for parameter. The first work assuming buildings to be represented as polyhedra is [2, 23]. Explicitely deriving neighbourhood relations between building parts was addressed by [24], see Fig. 4.

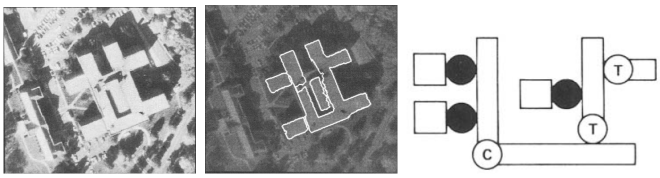

Fig. 4. Deriving the topology of a complex building. **Left:** Aerial image. **Mid:** Result of data driven segmentation. **Right:** Automatically derived symbolic image description; from [24]

Based on these stimulating results and motivated by the Avenches building extraction benchmark [25] we addressed the reconstruction of complex buildings from multiple images, see [26, 27]. Buildings are assumed to be hierarchically decomposed, to consist of building parts and its projection into the image yield corners, each being an aggregates of a point and its neighbouring edges and faces. The building parts are parametrized wire frames. The reconstruction method employees the integration of a data-driven trigger phase and a model-driven verification phase. In a first step, mutually oriented images 3D vertices were reconstructed (see Fig. 5), based on keypoints and neighbouring 3D edges in an prespecified area of interest, see [28, 29]. Based on the 3D vertices, building parts were hypothesized and mutual topological and geometrical constraints were exploited to reconstruct the

[1]The mutual self-information $I(x;y) \in (-\infty, \infty)$ depends on the probabilities. The mutual entropy $H(x;y) = \mathbb{E}_{p(x,y)}(I(x;y)) \geq 0$ is its expectation and often called mutual information.

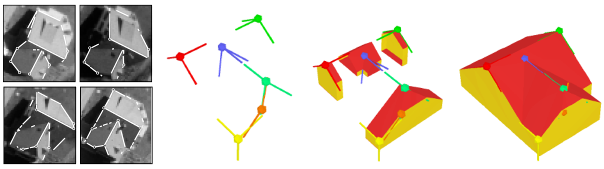

Fig. 5. 3D reconstruction of complex buildings from multiple images. **Left:** Four image sections. **Mid left:** Reconstructed 3D corners. **Mid right:** Triggered building parts: five terminals, one connector with three faces (belonging to the blue junction). **Right:** Reconstructed building (roof) fulfilling topological and geometrical constraints; from [27]

complete building. The method was implemented as constraint satisfaction program (in constraint logic programming, see [30, 31]), and allowed for occlusions (wild cards, see above) and incrementally for the prediction of new corners. The verification step included a prior on the different buildings and viewing directions, which exploited the shortest coding of the expected feature adjacency graph, see [32, 33].

Generating highly structured city models requires a quite generic building model, with a variable number of parts. The models we used are limited. They use restricted prespecified parametrized building parts, and thus cannot be used for larger areas. Though constraint logic programming appeared to be useful, the statistical knowledge only influenced the heuristics of the search and in a prespecified manner was used in the final evaluation. The parts need to be learned together with their relations and the reconstruction should exploit the learned statistics: Neither was the likelihood of the extracted features exploited as e.g., in [34], nor was any knowledge about illumination (sun angle, albedo) used. This would allow an integration of forward and backward modelling using computer graphics, see the example in Fig. 3.

There is a dichotomy: whether it is more favourable to aim at less simple parts with complex relations or to try to find more expressive complex parts with more restricted relations, with the inherent question how to deal with curved surfaces then. The extreme, representing the surface as mesh up to now appears to be the most flexible and successful approach. This circumvents the problem of structuring, which then needs to be addressed in a second step, see the next section.

Later work on building reconstruction exploited regularities of roof tops based on the straight skeleton [35] or aimed at watertight reconstruction for outdoor [36] or indoor scenes from LiDAR data [37]. We can observe intensive research in reconstructing large city areas based on terrestrial and aerial images using classical pipelines, which are developed in the context of reconstructing scenes from publically available images. This research is motivated by the difficulty in defining building parts as basic units, which are useful for larger areas, the high costs for directly acquiring terrestrial and aerial LiDAR data, and the need to provide textured scenes and hence avoids semantic structural descriptions; for pose estimation techniques for very large number of images see [38, 39]; for dense surface reconstruction see [40, 41, 42]. 3D surfaces of high fidelity and sufficient density, however, are an ideal basis for deriving semantically rich building descriptions, the topic of the next section.

C. *Image Interpretation with Graphical Models*

Deriving maps from images (including range images, e.g., LiDAR measurements) – a central task of photogrammetric research – by means of automatic image interpretation techniques still is in a premature state.

We addressed the problem of image interpretation for generating structured scene descriptions using building façades as exemplary domain. Façades show a wide variety in parts (doors, windows, balconies), structure (repetitions, symmetry, alignment) and appearance (local shadows, reflections, vegetation). Due to their mostly two-dimensional character modelling regularities is simpler than when dealing with general 3D building structures. We investigated two approaches: data driven semantic image segmentation using graphical models, especially conditional random fields (CRFs) [43, 44], and model driven façade reconstruction using marked point processes (MPPs) [45, 46]. We only discuss the model-driven approach.

The model driven reconstruction [45, 46] starts from rectified images, assuming the scale to be known. The model is a marked point process where façades consist of façade elements (doors, windows, balconies) represented as rectangles. The interpretation uses a reversible jump Markov chain Monte Carlo (rjMCMC) hypothesis and test paradigm. The geometric properties of the elements and their spatial relations are learned from training data. The data term of the energy function depends on a probabilistic object related classification; see Fig. 6. For the bottom example, observe the wrong heights of the windows, the confusion of windows and balconies and the detection of windows, where the ground truth does not indicate them; these errors can be explained by a too weak prior on the neighbourhood relations and the lack of long range interactions between the façade elements. The empirical evaluation of the

Fig. 6. Façade reconstruction based on a marked point process for façade elements (image, ground truth, reconstruction), CPU-time appr. one hour. from [45]

method leads to confusion tables, which contain estimated conditional probabilities, for which confidence intervals can be given. In order to arrive at reasonable intervals in case the empirical probability is 0 or 1, a weak Dirichlet prior for the multinomial distribution of these empirical probabilities can be used, see Table I.

TABLE I

Top: Confusion matrix for the façade type 'city houses' (with classes, background, window and balcony); **Bottom:** Corresponding 99%-confidence regions in % for the probabilities using a weak Dirichlet prior $\mathcal{D}(\boldsymbol{\alpha})$ with, e.g., $\boldsymbol{\alpha} = [1, 0.01, 0.01]$ for the first row. This yields more reasonable intervals for cases where the empirical probability is 0 or 1; from [45]

		prediction		
		bg	win	balc
truth	bg	0	**5**	0
	win	6	**146**	0
	balc	0	3	**40**

background	window	balcony
0.11 < 16.7 < *65.3*	*34.5* < **83.1** < *99.9*	*0.00* < 0.17 < *9.56*
1.03 < 3.93 < *9.04*	*90.9* < **96.1** < *99.0*	*0.00* < 0.01 < *0.36*
0.00 < 0.02 < *1.27*	*0.81* < 6.84 < *19.8*	*80.1* < **93.1** < *99.2*

We observed the typical strengths and weaknesses of data driven and model driven methods. *Data driven methods* are fast, can adapt locally to the image information and are versatile. This refers not only to locally connected Markov random fields (MRF), which, latest since grab-cut [47], pushed research in semantic segmentation, see the review [48]. This also holds for (1) fully connected MRFs, e.g., [49, 50, 51], which, due to their special assumption on the potentials, easily achieve real time [52] while still being competitive, (2) for autocontext models, which aim at sequentially gathering new context by using features of previous interpretations, see e.g., [53, 54, 55], but even more also (3) for convolutional neural networks, e.g., [56, 57, 58]. The techniques have also been applied successfully to semantically segmenting point clouds, e.g., [59, 60, 61]. Graphical models may be linked to logical programming via Markov logical networks [62]. They allow for a mixture of crisp and soft formulas [63]. They are used for event and face recognition in image sequence analysis [64, 65], for text understanding [66], for the interpretation of images of chemical structures [67], and for scene interpretation [68].

Model driven methods allow to explicitly model long range constraints. This in a first place holds for models based on *grammars* [69] and *marked point processes* [70]. Grammars are regularly for city modeling [71, 72, 73, 74, 75] or for roof extraction [76]. They allow learning, as for indoor scenes [74], for façades [77, 78, 79], for building layouts [80], or for architectural styles [81]. In image interpretation *marked point processes* are used for building extraction [70], for road network extraction [82], or more geometric feature extraction [83]. The generality of these models requires costly sampling methods for (approximately) finding optimal interpretations, which, however, allow for parallelization [84].

The *integration of data and model driven methods* has always been the key to successful interpretations. Early approaches, such as [85], used perceptual grouping techniques for providing candidate regions for object detection, here detecting buildings in aerial images. The same flavour can be found in recent work on simultaneous segmentation and detection [86], where the region proposals are refined after classification in order to obtain more accurate region boundaries.

A probably first integration of Markov logical networks and stochastical grammars for interpreting façades from point clouds is described in [87], see Fig. 7. The partonomy of

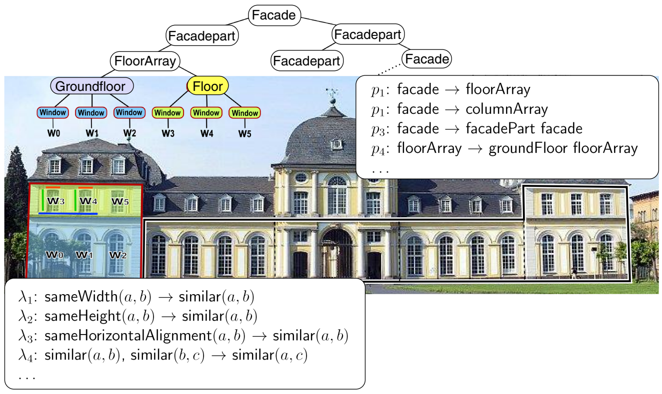

Fig. 7. Façade model with stochastical grammar and Markov logical network. **Upper left:** An instance of the grammar. **Upper right:** Some probabilistic rules of the grammar. **Lower left:** Some probabilistic relations of the Markov logic network; from [87]

Fig. 8. Interpretation of point clouds of façades with stochastic grammars and Markov logical networks. **Top row:** Image of façade. **Second row:** Point cloud with holes. **Third row:** Data driven interpretation of point cloud; windows: green), doors: pink, façade: white. **Last row:** Model driven interpretation. Observe the predictive power of the model; from [87]

the façade is represented in stochastic attributed grammatical rules, which capture the geometric properties and relations between the parts, see Fig. 7, upper right. Additional constraints are represented as predicates, which due to the diversity of the training data are give a probability. Relations between these predicates establish the Markov logic network, see Fig. 7, lower left. The interpretation of the point cloud starts with detecting basic parts of the façade. Deficiencies such as missing parts, or wrong alignments are then corrected using the prior mode, see Fig. 8.

III. Some Current Problems

This section discusses a few problems which regularly appear when developing methods for automatic interpretation of man made scenes. They address the choices we have when modelling the imaging process with the goal to solve the inverse problem, namely to recover scene information from images. Specifically, they refer to the model of the image signal, the relation between discrete and continuous geometry, the integration of crisp and soft prior knowledge, and the type of uncertainty of events and their meaning.

A. Physical and Phenomenological Signal Models

The basic steps for image orientation and building reconstruction, as the examples showed, often use methods for edge and contour detection, which essentially depend on the assumed image model. A classical model for the observed intensities $\boldsymbol{g}(i)$ in an image starts from the photon counts $\boldsymbol{N}(i)$ at each pixel in k channels: the two k-vectors $\boldsymbol{g}(i) \propto \boldsymbol{N}(i)$. This basic assumption leads to several problems, when following classical image processing procedures:

- How to exploit colour theory for non-RGB imagery? Colour theory models are a phenomenological and model visual perception of colours and its peculiarities, such as colour definition or colour constancy, or it models colour printing. For the majority of images available it may be useful: however, the analysis of images with more than three colours, even of hyperspectral images, the basic physical model appears to be the appropriate start. Improvements of classifiers using other than the original RGB signal result from reduced correlations, which are preferred by models which treat features as uncorrelated. Models, which take the – in principle arbitrary – distribution of the three colours into account, would not gain from colour transformations.
- Image intensities, being proportional to photon counts, are positive values. Representing a spatial intensity $g(i)$ as a sum of basis functions which are not non-negative, as when applying Fourier or Wavelet analysis, appears to be physically meaningless. Nevertheless, spectral methods have shown to be very successful.
- Since perception is logarithmic, a simple way out would be to work with the logarithms of the the intensities, as proposed by [88], who motivates it by the logarithmic perception of intensity.[2] The representation of the positive function then would be similar to the exponential family of densities, see [89] and the generalisation in [90], e.g., when assuming a continuous image domain, $f_l(x) = \log(g(x)) = \int G(u) \exp(2\pi i x u) \mathrm{d}u$, where $G(u)$ is the Fourier transform of $g(x)$.
- The statistical model of the observed intensities, being proportional to the photon counts, is a Poisson distribution. Then the variance of the intensity increases linearly with the intensity, omitting thermal noise and non-linearities of the sensor. Using a simple box-filter for smoothing implicitly assumes the intensities to have the same variance in the chosen neighbourhood, which does not hold. Checking the gradient magnitude for detecting edges, which is a classification task, should take the variance, i.e., the intensity level into account. Alternatively, the signal could be *variance normalized*, in the most simple case using a square root point transformation $f_s(x) = \sqrt{g(x)}$, since the normalized signal $f_s(x)$ then has constant noise variance, see [91, 92]. Many algorithms for keypoint detection could gain from such a transformation, leading to less keypoints in bright and more keypoints in dark areas of the image. This type of transformation also is motivated by the sensitivity of visual perception to image coding, see [93].

It would be desirable to have an integrating model for intensity signals in order to allow for efficient statistical, physical and (spatial) spectral analysis. A scale analysis of the factors resulting from non-negative signal factorization may play a guiding role.

[2] [88] in addition allow for affine transformations of the logarithm of the intensity.

B. Discrete and Continuous Geometry

Recovering man-made objects aims at some geometric description of the object's boundary, which usually is represented as an aggregation of continuous surface regions in 3D. The expected image of such a piecewise surface is a partitioning of the image region with piecewise boundaries, where not all intensity edges necessarily need to have two distinct regions as neighbours. The observed image grid as observed 3D point clouds are discrete. Hence, the reconstruction of the continuous 3D surface regions and their boundaries consists (1) of the topologically consistent identification of these boundaries and (2) the geometrically consistent determination of the form of the surfaces and the boundaries. An example where boundaries may not be detectable due to lighting conditions and may lead to violations of the image model as is given in Fig. 9.

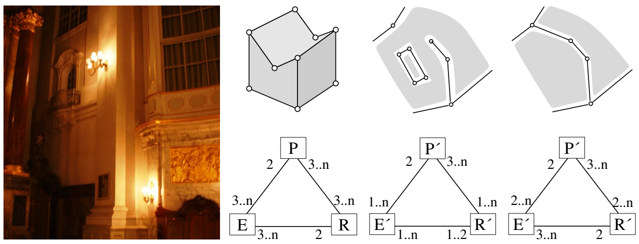

Fig. 9. Topological relations for polyhedra and their ideal and real (extracted) images. **Left:** Image of a vertical edge appearing with zero gradient (St. Michaelis church, Hamburg). The following: Example image and entity relation diagram with range of multiplicities. **Mid left:** Polyhedral boundary; edges (E) may must have two neighbouring regions (R) in 3D. **Mid:** Ideal image of polyhedral boundary, admitting zero gradient edges; edges may have 1 to 2 neighbouring regions in the ideal image. **Right:** Real image from partitioning; edges (E) must have two neighbouring regions (R) in the partitioning. In all cases edges have two neighbouring (end) points (P). Obviously, the ideal image does **not** follow the winged-edge representation

The recovery of a consistent boundary description is underconstrained, unless the point density (of the grid or the point cloud) is sufficiently high and the boundary lines fulfil certain regularities, see e.g., the sampling theorem for recovering region boundaries [94], architectural models [95], not necessarily based on triangular meshes [96, 97].

Moreover, grid-based methods, such as Markov random fields, do not allow to include prior knowledge about the straightness of boundaries. Therefore, algorithms for finding consistent polygonal boundaries mostly contain ad hoc rules, cannot include statistical prior information about the observed points, and – due to the occurrence of structural errors – are difficult to be evaluated.

This touches the integration of bottom-up and top-down procedures, discussed above: geometric entities, such as polygons or polygon networks, being mid-level structures, require a statistically coherent modelling of both, their appearance – for bottom-up hypothesis building – as well as their geometric and neighbourhood relations – for top-down prediction; this appears like bi-directional search in the solution space. Reducing the costs for sampling from large energy models, such as with MCMC, is described in [98].

C. Crisp and Soft Prior Knowledge

Handling both, crisp and soft prior knowledge, as prior is essential (not only) for interpreting images of man-made scenes. Taxonomies and partonomies of objects and spatial relations, such as parallelity, play a central role in semantic modelling, see the discussion of the role of semantics for games in [99]. The uncertainty of observations and models and the success of probabilistic models is ubiquitous.

It is less clear how those parts of the model, which are certain (in a probabilistic sense), are handled in a principled manner: i.e., explicitly. *Geometric relations* in multi-view analysis usually are hard coded; algebraic methods, such as Gröbner bases, though a research topic on its own, increasingly are used to derive solutions, but are not integrated into systems, where the task is not fixed. Attempts to use algebraic methods for more generic tasks, have been intensively discussed in the late eighties, see [100, 101]. Methods which detect regularities and use them for the enhancement of 2D and 3D objects, such as in [102, 103], have to face the the inconsistency of individual hypothesis tests or the explosion of computational complexity – prior to finding bases for the constraints, which then can be applied.

Partonomies and taxonomies are increasingly used for improving categorization [104, 105]. Following [106], the simultaneous classification of a category and a subcategory is significantly better than the individual classification. Explicitly classifying image galleries, i.e., ensembles of images, into a given taxonomy (derived from Wikipedia) is adressed by [107]. The images in the data base IMAGENET [108] are organized in a semantic hierarchy (WordNet), supporting benchmarking of classifiers which can exploit this knowledge. Since ImageNet is based on the ontology of WordNet it would be desirable to have the concepts around 'building' for interpreting outdoor and indoor images included in ImageNet. Since WordNet is focussed on function of notions and does not include any concepts for geometric or material the link between semantic, geometric, and radiometric models still remains to be established, e.g., for the domains 'building' and 'road', possibly exploiting grammars, marked point processes, or Markov logic networks, see the example above.

In this context two questions arise. First, what are the adequate methods to learn the models, i.e., the geometric and semantic relations? Learning the structure and the parameters of probabilistic logic, where clauses are attached with a probability, may be based on measuring the success of data base queries [109, 63]. Learning structures can use the development in kernel methods, which allow to address all types of structures: multi-label, with taxonomies, label-sequence-learning, sequence of operations alignment, natural language parsing, see [110].

Second, what are efficient interpretation processes? There exist several methods to derive statistically interpretations based on crisp and uncertain information, e.g., using probabilistic logic programming, statistical relational learning, or Markov logic, see the overview in [63] and the Dagstuhl Seminar on *Logic and Probability for Scene Interpretation*; see [111]. Attempts to increase efficiency use a reduced language e.g., [112], or apply sampling techniques, e.g., [113, 114]. Except for a few examples, e.g., [115, 87], see above, the techniques are not yet exploited for analysing images, especially of man-made objects.

D. Uncertainty and Vagueness

Decision making using classifiers always has assumed that data, models, and decisions are uncertain. However, the process and the result of classifiers often do not reflect this uncertainty.

First, many classifiers only report the most likely class for each object in a 'winner takes all' habit. This does not support the need of a user to know the uncertainty of the decision. Even giving a confusion matrix, often is not sufficient, as the estimated conditional probabilities are estimates, and hence are uncertain. Giving confidence regions as in Table I, would be a first remedy. Results of [116, 117] indicate, that import vector machines [118] yield more reliable posterior probabilities than the output of support vector machines, when transforming their output into probabilities [119]. Since the output of classifiers often is used for generating the potentials of Markov random fields, their quality may have a decisive impact.

The uncertainty of semantic segmentation cannot be represented with confusion tables, as the space of segmentations is far too large, why indicating the uncertainty of the boundaries appears a reasonable approach, see e.g., [120, 121] both using deep convolutional network.

Second, many classifiers assume that each object belongs to one of the presumed classes, possibly a rejection class. This has been found to be over-simplistic. Images with complex content may belong to different classes, e.g., a natural scene may simultaneously be classified as mountain area and beach area, if ingredients (key features) for both classes can be detected, and the designer of the classifier intends such an overlap of classes, see [34, 122] and the review [123].

Third, the classes themselves are difficult to separate, e.g., the two classes low vegetation and high vegetation in the 'Large Scale Point Cloud Classification benchmark'.[3] This type of uncertainty in the definition of classes was the motivation for developing fuzzy models [124, 125], where each object may belong to a class according to some membership value. The heavy debate on the relation between fuzzy theory and probability theory is reflected and resolved in the key paper by Dubois and Prade [126]: The semantic distinction between the vagueness/fuzziness of the notion of an event and the uncertainty/likelihood of the existence or appearance of an event indicates the two notions to be orthogonal; integrating both concepts, while keeping their key properties, such as [127, 125, 128], still seems to have no canonical solution.

Anyhow, when taking into account the necessity to handle non-unique ground truth in benchmarks [129], to deal with occlusions,[4], and to take vaguely defined classes into account, when evaluating classifiers (see [130]), then the number of papers addressing fuzzy logic on international conferences such as ICPR, ECCV, and ICCV, being below 0.5 % on an average, appears to be very low.

IV. A Future for Learning Building Models

The paper has addressed various aspects of interpreting images of man-made structures, especially of buildings. It focused on methods, which reflect the underlying models of the imaging, analysis and interpretation processes and which hence allow the user of such a system to make decisions, thus to understand the image content in an appropriate manner. The problems, mentioned in the last section, all are caused by insufficiencies or incompatibilities of simultaneously applied models. The tools to solve or overcome these problems appear to be available.

We discussed the dichotomy of discriminative and generative models, both having their advantages. Generative model are efficient in obtaining quantitatively good results, while generative models are powerful in elucidating structured semantics. The dichotomy is best seen in semantic segmentation: The partitioning of image into relevant regions requires a process which is at the same time data and model driven. This motivates the structuring of the interpretation/understanding task as in Fig. 10.

All processes can gain from the interaction of recognition, reconstruction and re-organization, proposed in [22]. Generative models also are directly amenable to incremental learning. On the other hand the speed of current neural network classification and regression tools, which does in no way correspond to the generally long training times, contrasts to the fast training times of explicit semantic models, such as grammars of marked point processes, which are often much slower in reasoning. Attempts to use neural network priors for

[3]See http://www.semantic3d.net

[4]See e.g., the annotation rules in the PASCAL http://host.robots.ox.ac.uk/pascal/VOC/voc2008/guidelines.html.

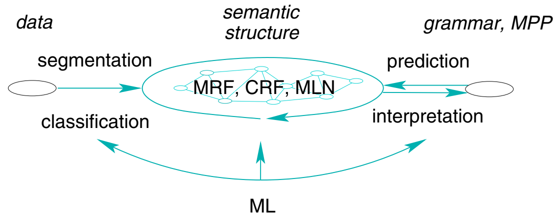

Fig. 10. Metamodel for interpretion and understanding of image data (left ellipse). Preknowledge (right ellipse), e.g., in the form of grammars or marked point processes (MPP), are necessary, in order to capture the envisaged meaning of the interpretation. Intermediate structures (mid ellipse) may be represented and analysed e.g., by Markov random fields (MRF), conditional random fields (CRF) or Markov logic networks (MLN). These structures are *simultaneously* predicted from the preknowledge and from the data by – possibly semantic – segmentation. The parameters of all processes (indicated with upright letters and arrows) are trained using techniques from machine learning (ML). The control of the complete process is an open problem

one-shot learning, such as [131], are promising. The flexibility of multi-layer neural networks also needs to be compared with the rich representation of the scattering transform [132, 133], which code higher moments of the underlying signal, and have been applied in face recognition [134], used for graphs [135], and enriched by rotation invariant kernels [136].

The trend to have very large and rich bodies of image data for benchmarking can be interpreted as extensionally defining what an image is, instead of intentionally modelling images by power spectra, higher order characteristics, or random fields.

A future for learning highly structured models may be based the available basic technology which not yet is exploited for establishing rich geometric and semantic building models.

References

[1] R. Brooks, "Model-based 3-D interpretation of 2-D images," *IEEE T-PAMI*, vol. 5, no. 2, pp. 140–150, 1983.

[2] M. Herman and T. Kanade, "Incremental Reconstruction of 3D Scenes from Multiple, Complex Images," *Artificial Intelligence*, vol. 30, pp. 289–341, 1986.

[3] P. Musialski, P. Wonka, D. G. Aliaga, M. Wimmer, L. van Gool, and W. Purgathofer, "A Survey of Urban Reconstruction," *Computer Graphics Forum*, vol. 32, no. 6, pp. 146–177, Sep. 2013. [Online]. Available: http://www.cg.tuwien.ac.at/research/publications/2012/musialski-2012-surcgf/http://doi.wiley.com/10.1111/cgf.12077

[4] S. Fidler and R. Urtasun. (2015) 3D Indoor Scene reconstruction. http://www.cs.toronto.edu/~fidler/slides/CVPR15tutorial/. [Online]. Available: http://www.cs.toronto.edu/~fidler/slides/CVPR15tutorial/

[5] B. M. Lake, T. D. Ullman, J. B. Tenenbaum, and S. J. Gershman, "Building Machines That Learn and Think Like People," *CoRR*, vol. abs/1604.00289, 2016. [Online]. Available: http://arxiv.org/abs/1604.00289

[6] R. A. Brooks, R. Creiner, and T. O. Binford, "The ACRONYM Model-based Vision System," in *Proceedings of the 6th International Joint Conference on Artificial Intelligence - Volume 1*, ser. IJCAI'79. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 1979, pp. 105–113. [Online]. Available: http://dl.acm.org/citation.cfm?id=1624861.1624890

[7] D. G. Lowe, "Three-Dimensional Object Recognition from Single Two-Dimensional Images," *Artificial Intelligence*, vol. 31, pp. 355–395, 1987.

[8] W. E. L. Grimson and T. Lozano-Perez, "Localizing Overlapping Parts by Searching the Interpretation Tree," *IEEE T-PAMI*, vol. 9, no. 4, pp. 469–482, 1987.

[9] R. M. Haralick and L. G. Shapiro, "The Consistent Labeling Problem: Part 1," *IEEE T-PAMI*, vol. 1, no. 2, pp. 173–184, 1979.

[10] ——, "The Consistent Labeling Problem: Part 2," *IEEE T-PAMI*, vol. 2, no. 3, pp. 193–203, 1980.

[11] L. G. Shapiro and R. M. Haralick, "Relational Matching," *Applied Optics*, vol. 26, no. 10, pp. 1845–1851, may 1987.

[12] W. Förstner, "Model Based Detection and Location of Houses as Topographic Control Points in Digital Images," in *Intl. Archives of Photogrammetry and Remote Sensing*, vol. 27, no. Part B11. XVIth ISPRS Congress, Kyoto, 1988.

[13] W. Schickler, "Feature Matching for Outer Orientation of Single Images Using 3-D Wireframe Controlpoints," in *Intl. Archives of Photogrammetry and Remote Sensing*, vol. 29, no. Part B3. Proc. XVIIth ISPRS Congress, Washington, D. C., 1992, pp. 591–598.

[14] G. Vosselman, *Relational Matching*, ser. Lecture Notes on Computer Science. Springer Verlag, 1992, vol. 628.

[15] N. Haala and G. Vosselman, "Recognition of Road and River Patterns by Relational Matching," in *Intl. Archives for Photogrammetry*, vol. XXIX, Part B3, 1992, pp. 969–975.

[16] G. Vosselman and N. Haala, "Erkennung topographischer Paßpunkte durch relationale Zuordnung," *Zeitschrift für Photogrammetrie und Fernerkundung*, vol. 6, pp. 170–176, 1992.

[17] E. Brachmann, F. Michel, A. Krull, M. Y. Yang, S. Gumhold, and C. Rother, "Uncertainty-Driven 6D Pose Estimation of Objects and Scenes from a Single RGB Image," in *Proc. of Conf. on Computer Vision and Pattern Recognition*, 2016.

[18] S. Tulsiani and J. Malik, "Viewpoints and Keypoints," in *Proc. of Conf. on Computer Vision and Pattern Recognition*. IEEE, 2015, pp. 1510–1519.

[19] R. Mottaghi, Y. Xiang, and S. Savarese, "A Coarse-to-Fine Model for 3D Pose Estimation and Sub-category Recognition," *CoRR*, vol. abs/1504.02764, 2015. [Online]. Available: http://arxiv.org/abs/1504.02764

[20] A. Kar, S. Tulsiani, J. Carreira, and J. Malik, "Category-Specific Object Reconstruction from a Single Image," *CoRR*, vol. abs/1411.6069, 2014. [Online]. Available: http://arxiv.org/abs/1411.6069

[21] Y. Xiang, R. Mottaghi, and S. Savarese, "Beyond PASCAL: A Benchmark for 3D Object Detection in the Wild," in *IEEE Winter Conference on Applications of Computer Vision (WACV)*, 2014.

[22] J. Malik, P. Arbelez, J. Carreira, K. Fragkiadaki, R. Girshick, G. Gkioxari, S. Gupta, B. Hariharan, A. Kar, and S. Tulsiani, "The Three R's of Computer Vision: Recognition, Reconstruction and Reorganization," *Pattern Recognition Letters*, vol. 72, pp. 4–14, 2015.

[23] A. Huertas and R. Nevatia, "Detecting Buildings in Aerial Images," *Computer Vision, Graphics, and Image Processing*, no. 41, pp. 131–152, 1988.

[24] P. Fua and A. J. Hanson, "Resegmentation Using Generic Shape Locating General Cultural Objects," *Pattern Recognition Letters*, vol. 5, pp. 243–252, 1987.

[25] S. Mason, M. Baltsavias, and D. Stallmann, "High Precision Photogrammetric Data Set for Building Reconstruction and Terrain Modelling," 1994, eTH Zürich. [Online]. Available: www.isprs.org/data/avenches/dataset_descr.ps

[26] C. Braun, T. H. Kolbe, F. Lang, W. Schickler, V. Steinhage, A. Cremers, W. Förstner, and L. Plümer, "Models for Photogrammetric Building Reconstruction," *Computer & Graphics*, vol. 19, no. 1, pp. 109–118, 1995.

[27] A. Fischer, T. H. Kolbe, F. Lang, A. B. Cremers, W. Förstner, L. Plümer, and V. Steinhage, "Extracting Buildings from Aerial Images Using Hierarchical Aggregation in 2D and 3D," *CVIU*, vol. 72, no. 2, pp. 185–203, 1998.

[28] C. Fuchs and W. Förstner, "Polymorphic Grouping for Image Segmentation," in *Proc. 5th ICCV*. IEEE Computer Society Press, 1995, pp. 175–182.

[29] F. Lang and W. Förstner, "Surface Reconstruction of Man-Made Objects using Polymorphic Mid-Level Features and Generic Scene Knowledge," *Zeitschrift für Photogrammetrie und Fernerkundung*, vol. 6, pp. 193–201, 1996.

[30] T. H. Kolbe, L. Plümer, and A. B. Cremers, "Identifying Buildings in Aerial Images Using Constraint Relaxation and Variable Elimination," *IEEE Intelligent Systems*, vol. 15, no. 1, pp. 33–39, Jan. 2000. [Online]. Available: http://dx.doi.org/10.1109/5254.820327

[31] A. Fischer, "Automatische Gebäuderekonstruktion mittels parametrisierter Komponenten," Ph.D. dissertation, Insititut für Informatik, Universität Bonn, eingereicht, 2000.

[32] S. Heuel and T. H. Kolbe, "Building Reconstruction: The Dilemma of Generic Versus Specific Models," *Künstliche Intelligenz*, vol. 3, pp. 57–62, July 2001. [Online]. Available: http://www.ipb.uni-bonn.de/ipb/lit/abstracts01/heuel01.building.html

[33] T. H. Kolbe, *Identifikation und Rekonstruktion von Gebäuden in Luftbildern mittels unscharfer Constraints*. Shaker Verlag: Shaker, 2000.

[34] P. Arbelaez, M. Maire, C. Fowlkes, and J. Malik, "Contour Detection and Hierarchical Image Segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 33, no. 5, pp. 898–916, May 2011. [Online]. Available: http://dx.doi.org/10.1109/TPAMI.2010.161

[35] C. Brenner, "Towards Fully Automatic Generation of City Models," in *International Archives of Photogrammetry and Remote Sensing*, vol. XXXIII, no. B3, 2000, pp. 85–92.

[36] Q.-Y. Zhou and U. Neumann, "2.5D Dual Contouring: A Robust Approach to Creating Building Models from Aerial LiDAR Point Clouds," in *Proceedings of the 11th European Conference on Computer Vision Conference on Computer Vision: Part III*, ser. ECCV'10. Berlin, Heidelberg: Springer-Verlag, 2010, pp. 115–128. [Online]. Available: http://dl.acm.org/citation.cfm?id=1927006.1927017

[37] S. Oesau, F. Lafarge, and P. Alliez, "Indoor Scene Reconstruction using Primitive-driven Space Partitioning and Graph-cut," in *Eurographics Workshop on Urban Data Modelling and Visualisation*, Girona, Spain, May 2013. [Online]. Available: https://hal.inria.fr/hal-00814546

[38] N. Snavely, S. M. Seitz, and R. Szeliski, "Photo Tourism: Exploring Photo Collections in 3D," *ACM Trans. Graph.*, vol. 25, no. 3, pp. 835–846, Jul. 2006. [Online]. Available: http://doi.acm.org/10.1145/1141911.1141964

[39] J.-M. Frahm, P. Fite-Georgel, D. Gallup, T. Johnson, R. Raguram, C. Wu, Y.-H. Jen, E. Dunn, B. Clipp, S. Lazebnik, and M. Pollefeys, "Building Rome on a Cloudless Day," in *Proceedings of the 11th European Conference on Computer Vision: Part IV*, ser. ECCV'10. Berlin, Heidelberg: Springer-Verlag, 2010, pp. 368–381. [Online]. Available: http://dl.acm.org/citation.cfm?id=1888089.1888117

[40] Y. Furukawa and J. Ponce, "Accurate, Dense, and Robust Multiview Stereopsis," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 32, no. 8, pp. 1362–1376, Aug. 2010. [Online]. Available: http://dx.doi.org/10.1109/TPAMI.2009.161

[41] M. Jancosek and T. Pajdla, "Multi-view Reconstruction Preserving Weakly-supported Surfaces," in *Proc. of Conf. on Computer Vision and Pattern Recognition*, ser. CVPR '11. Washington, DC, USA: IEEE Computer Society, 2011, pp. 3121–3128. [Online]. Available: http://dx.doi.org/10.1109/CVPR.2011.5995693

[42] F. Langguth, K. Sunkavalli, and M. G. Sunil Hada and, "Shading-aware Multi-view Stereo," in *Proc. of European*

Conference on Computer Vision, 2016.

[43] F. Korč, "Tractable Learning for a Class of Global Discriminative Models for Context Sensitive Image Interpretation," Ph.D. dissertation, Department of Photogrammetry, University of Bonn, http://hss.ulb.uni-bonn.de/2012/3010/3010.htm, 2012. [Online]. Available: http://hss.ulb.uni-bonn.de/2012/3010/3010.htm

[44] M. Y. Yang, "Hierarchical and Spatial Structures for Interpreting Images of Man-made Scenes Using Graphical Models," Ph.D. dissertation, Institute of Photogrammetry, University of Bonn, http://hss.ulb.uni-bonn.de/2012/2765/2765.htm, 2011. [Online]. Available: http://hss.ulb.uni-bonn.de/2012/2765/2765.htm

[45] S. Wenzel, "High-level facade image interpretation using marked point processes," Ph.D. dissertation, Department of Photogrammetry, University of Bonn, 2016. [Online]. Available: http://hss.ulb.uni-bonn.de/2016/4412/4412.htm

[46] S. Wenzel and W. Förstner, "Facade Interpretation Using a Marked Point Process," in *Int. Ann. Photogramm. Remote Sens. (ISPRS'16)*, 2016. [Online]. Available: http://www.ipb.uni-bonn.de/pdfs/Wenzel2016Facade.pdf

[47] C. Rother, V. Kolmogorov, and A. Blake, "GrabCut: Interactive Foreground Extraction using Iterated Graph Cuts," *ACM Transactions on Graphics*, vol. 23, pp. 309–314, 2004.

[48] H. Zhu, F. Meng, J. Cai, and S. Lu, "Beyond Pixels: A Comprehensive Survey from Bottom-up to Semantic Image Segmentation and Cosegmentation," *CoRR*, vol. abs/1502.00717, 2015. [Online]. Available: http://arxiv.org/abs/1502.00717

[49] P. Krähenbühl and V. Koltun, "Efficient Inference in Fully Connected CRFs with Gaussian Edge Potentials," in *Advances in Neural Information Processing Systems 24*, J. Shawe-Taylor, R. S. Zemel, P. L. Bartlett, F. Pereira, and K. Q. Weinberger, Eds. Curran Associates, Inc., 2011, pp. 109–117. [Online]. Available: http://books.nips.cc/papers/files/nips24/NIPS2011_0097.pdf

[50] K. Ristovski, V. Radosavljevic, S. Vucetic, and Z. Obradovic, "Continuous Conditional Random Fields for Efficient Regression in Large Fully Connected Graphs," in *Proceedings of the Twenty-Seventh AAAI Conference on Artificial Intelligence*, 2013.

[51] W. Li and M. Y. Yang, "Efficient semantic segmentation of man-made scenes using fully-connected conditional random fields," in *Archives of ISPRS*, 2016.

[52] M.-M. Cheng, V. A. Prisacariu, S. Zheng, P. H. S. Torr, and C. Rother, "DenseCut: Densely Connected CRFs for Realtime GrabCut," *Computer Graphics Forum*, vol. 34, no. 7, pp. –, 2015.

[53] Z. Tu, "Auto-context and Its Application to High-level Vision Tasks," in *Proc. of Conf. on Computer Vision and Pattern Recognition*, 2008.

[54] V. Jampani, R. Gadde, and P. V. Gehler, "Efficient facade segmentation using auto-context," in *2015 IEEE Winter Conference on Applications of Computer Vision, WACV 2015, Waikoloa, HI, USA, January 5-9, 2015*, 2015, pp. 1038–1045. [Online]. Available: http://dx.doi.org/10.1109/WACV.2015.143

[55] R. Gadde, V. Jampani, R. Marlet, and P. V. Gehler, "Efficient 2D and 3D Facade Segmentation using Auto-Context," *CoRR*, vol. abs/1606.06437, 2016. [Online]. Available: http://arxiv.org/abs/1606.06437

[56] C. Farabet, C. Couprie, L. Najman, and Y. LeCun, "Learning Hierarchical Features for Scene Labeling," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, August 2013.

[57] J. Long, E. Shelhamer, and T. Darrell, "Fully Convolutional Networks for Semantic Segmentation," *CoRR*, vol. abs/1411.4038, 2014. [Online]. Available: http://arxiv.org/abs/1411.4038

[58] D. Marmanisa, J. D. Wegner, K. S. S. Gallian and, M. Datcu, and U. Stilla, "Semantic segmentation of aerial images with an essemble of CNNs," in *Archives of ISPRS*, 2016.

[59] A. Adan, X. Xiong, B. Akinci, and D. Huber, "Automatic Creation of Semantically Rich 3D Building Models from Laser Scanner Data," in *Proceedings of the International Symposium on Automation and Robotics in Construction (ISARC)*, June 2011.

[60] M. Tamke, I. Blümel, S. Ochmann, R. Vock, and R. Wessel, "From Point Clouds to Definitions of Architectural Space," in *Fusion - 32nd International Conference on Education and research in Computer Aided Architectural Design in Europe*, ser. eCAADe: Conferences, vol. 2, Northumbria University. Newcastle, UK: Northumbria University, 2014, pp. 557–566.

[61] S. Ochmann, R. Vock, R. Wessel, and R. Klein, "Automatic Reconstruction of Parametric Building Models from Indoor Point Clouds," *Computers & Graphics*, vol. 54, pp. 94–103, Feb. 2016, special Issue on CAD/Graphics 2015. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0097849315001119

[62] M. Richardson and P. Domingos, "Markov logic networks," *Machine Learning*, vol. 62, no. 1, pp. 107–136, 2006. [Online]. Available: http://dx.doi.org/10.1007/s10994-006-5833-1

[63] D. Fierens, G. Van Den Broeck, J. Renkens, D. Shterionov, B. Gutmann, I. Thon, G. Janssens, and L. de Raedt, "Inference and Learning in Probabilistic Logic Programs using Weighted Boolean Formulas," *Theory and Practice of Logic Programming*, vol. 15, no. 3, pp. 358–401, 2014.

[64] S. D. Tran and L. S. Davis, "Event modeling and recognition using markov logic networks," in *Proceedings of the 10th European Conference on Computer Vision: Part II*, ser. ECCV '08. Berlin, Heidelberg: Springer-Verlag, 2008, pp. 610–623. [Online]. Available: http://dx.doi.org/10.1007/978-3-540-88688-4_45

[65] A. Chechetka, D. Dash, and M. Philipose, "Relational Learning for Collective Classification of Entities in Images," in *Proceedings of the 6th AAAI Conference on Statistical Relational Artificial Intelligence*, ser. AAAIWS'10-06. AAAI Press, 2010, pp. 7–12. [Online]. Available: http://dl.acm.org/citation.cfm?id=2908567.2908569

[66] H. Poon, "Markov Logic for Machine Reading," Ph.D. dissertation, University of Washington, 2011.

[67] P. Frasconi, F. Gabbrielli, M. Lippi, and S. Marinai, "Markov Logic Networks for Optical Chemical Structure Recognition," *J. of Chemical Information and Modeling*, vol. 54, no. 8, pp. 2380–2390, 2014.

[68] M. Xu and M. Petrou, "Learning Logic Rules for Scene Interpretation Based on Markov Logic Networks," in *Proceedings of the 9th Asian Conference on Computer Vision - Volume Part III*, ser. ACCV'09. Berlin, Heidelberg: Springer-Verlag, 2010, pp. 341–350. [Online]. Available: http://dx.doi.org/10.1007/978-3-642-12297-2_33

[69] S.-C. Zhu and D. Mumford, "A Stochastic Grammar of Images," *Found. Trends. Comput. Graph. Vis.*, vol. 2, no. 4, pp. 259–362, Jan. 2006. [Online]. Available: http://dx.doi.org/10.1561/0600000018

[70] M. Ortner, X. Descombes, and J. Zerubia, "Building Outline Extraction from Digital Elevation Models Using Marked Point Processes," *Int. J. Comput. Vision*, vol. 72, no. 2, pp. 107–132, Apr. 2007. [Online]. Available: http://dx.doi.org/10.1007/s11263-005-5033-7

[71] A. R. Dick, P. H. S. Torr, and R. Cipolla, "A Bayesian Estimation of Building Shape using MCMC," in *Computer Vision–ECCV 2002*, ser. LNCS, vol. 2351. Springer, 2002, pp. 574–575.

[72] J. O. Talton, Y. Lou, S. Lesser, J. Duke, R. Měch, and V. Koltun, "Metropolis Procedural Modeling," *ACM Trans.*

Graph., vol. 30, no. 2, pp. 11:1–11:14, Apr. 2011. [Online]. Available: http://doi.acm.org/10.1145/1944846.1944851
[73] A. Martinović and L. V. Gool, "Bayesian Grammar Learning for Inverse Procedural Modeling," in *Proc. of Conf. on Computer Vision and Pattern Recognition*, 2013.
[74] T. Liu, S. Chaudhuri, V. G. Kim, Q. Huang, N. J. Mitra, and T. Funkhouser, "Creating Consistent Scene Graphs Using a Probabilistic Grammar," *ACM Trans. Graph.*, vol. 33, no. 6, pp. 211:1–211:12, Nov. 2014. [Online]. Available: http://doi.acm.org/10.1145/2661229.2661243
[75] M. Schwarz and P. Müller, "Advanced Procedural Modeling of Architecture," *ACM Trans. Graph.*, vol. 34, no. 4, pp. 107:1–107:12, Jul. 2015. [Online]. Available: http://doi.acm.org/10.1145/2766956
[76] H. Huang, C. Brenner, and M. Sester, "3D Building Roof Reconstruction from Point Clouds via Generative Models," in *Proceedings of the 19th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*, ser. GIS '11. New York, NY, USA: ACM, 2011, pp. 16–24. [Online]. Available: http://doi.acm.org/10.1145/2093973.2093977
[77] N. Ripperda and C. Brenner, "Evaluation of Structure Recognition using Labelled Facade Images," in *Pattern Recognition: 31st DAGM Symposium*, J. Denzler, G. Notni, and H. Süße, Eds. Springer, 2009, pp. 532–541.
[78] L. Fan, P. Musialski, L. Liu, and P. Wonka, "Structure Completion for Facade Layouts," *ACM Trans. Graph.*, vol. 33, no. 6, pp. 210:1–210:11, Nov. 2014. [Online]. Available: http://doi.acm.org/10.1145/2661229.2661265
[79] F. Wu, D.-M. Yan, W. Dong, X. Zhang, and P. Wonka, "Inverse Procedural Modeling of Facade Layouts," *ACM Trans. Graph.*, vol. 33, no. 4, pp. 121:1–121:10, Jul. 2014. [Online]. Available: http://doi.acm.org/10.1145/2601097.2601162
[80] F. Bao, D.-M. Yan, N. J. Mitra, and P. Wonka, "Generating and Exploring Good Building Layouts," *ACM Transactions on Graphics*, vol. 32.4, p. 122, 2013.
[81] J. O. Talton, L. Yang, R. Kumar, M. Lim, N. D. Goodman, and R. Měch, "Learning Design Patterns with Bayesian Grammar Induction," in *Proceedings of the 25th annual ACM symposium on user interface software and technology.*, 2012.
[82] D. Chai, W. Förstner, and F. Lafarge, "Recovering Line-Networks in Images by Junction-Point Processes," in *Conf. on Computer Vision and Pattern Recognition*, 2013.
[83] F. Lafarge, G. Gimel'farb, and X. Descombes, "Geometric Feature Extraction by a Multi-Marked Point Process," *Transaction on Pattern Analysis and Machine Intelligence*, vol. 32, pp. 1597–1609, 2008.
[84] D. J. Wilkinson, *Parallel bayesian computation*. Marcel Dekker/CRC Press, 2006, ch. 18, pp. 481–512.
[85] R. Mohan and R. Nevatia, "Using Perceptual Organization to extract 3D-structures," *IEEE T-PAMI*, vol. 11, pp. 1121–1139, 1989.
[86] B. Hariharan, P. Arbelaez, R. B. Girshick, and J. Malik, "Simultaneous Detection and Segmentation," *CoRR*, vol. abs/1407.1808, 2014. [Online]. Available: http://arxiv.org/abs/1407.1808
[87] Y. Dehbi, F. Hadiji, G. Gröger, K. Kersting, and L. Plümer, "Statistical Relational Learning of Grammar Rules for 3D Building Reconstruction," *Transactions in GIS*, 2016.
[88] J. J. Koenderink and A. J. v. Doorn, "Image Processing Done Right," in *Proceedings of the 7th European Conference on Computer Vision-Part I*, ser. ECCV '02. London, UK, UK: Springer-Verlag, 2002, pp. 158–172. [Online]. Available: http://dl.acm.org/citation.cfm?id=645315.649167
[89] J. M. Borwein and W. Z. Huang, "A Fast Heuristic Method for Polynomial Moment Problems with Boltzmann-Shannon Entropy," *SIAM Journal on Optimization*, vol. 5, no. 1, pp. 68–99, 1995.
[90] D. Fasino, "Approximation of nonnegative functions by means of exponentiated trigonometric polynomials," *J. Comput. Appl. Math.*, vol. 140, no. 1-2, pp. 315–329, Mar. 2002. [Online]. Available: http://dx.doi.org/10.1016/S0377-0427(01)00406-X
[91] W. Förstner, "Image Preprocessing for Feature Extraction in Digital Intensity, Color and Range Images," in *Geomatic Methods for the Analysis of Data in Earth Sciences*, ser. Lecture Notes in Earth Sciences. Springer, 2000, vol. 95/2000, pp. 165–189. [Online]. Available: http://www.ipb.uni-bonn.de/pdfs/Forstner2000Image.pdf
[92] B. Jähne and M. Schwarzbauer, "Noise equalisation and quasi loss-less image data compression – or how many bits needs an image sensor?" *Technisches Messen, De Gruyter*, vol. 83, pp. 16–24, 2016.
[93] J. L. Mannos and D. J. Sakrison, "The Effects of a Visual Fidelity Criterion on the Encoding of Images," *IEEE Transactions on Information Theory*, vol. IT-20, no. 4, pp. 525–536, July 1974.
[94] H. Meine, U. Köthe, and P. Stelldinger, "A Topological Sampling Theorem for Robust Boundary Reconstruction and Image Segmentation," *Discrete Appl. Math.*, vol. 157, no. 3, pp. 524–541, Feb. 2009. [Online]. Available: http://dx.doi.org/10.1016/j.dam.2008.05.031
[95] H. Pottmann, M. Eigensatz, A. Vaxman, and J. Wallner, "Architectural geometry," *Computers and Graphics*, vol. 47, pp. 145–164, 2015. [Online]. Available: http://www.geometrie.tuwien.ac.at/geom/fg4/sn/2015/ag/ag.pdf
[96] Y. Liu, H. Pottmann, J. Wallner, Y.-L. Yang, and W. Wang, "Geometric Modeling with Conical Meshes and Developable Surfaces," *ACM Trans. Graph.*, vol. 25, no. 3, pp. 681–689, Jul. 2006. [Online]. Available: http://doi.acm.org/10.1145/1141911.1141941
[97] D. Kovacs, A. Myles, and D. Zorin, "Anisotropic quadrangulation," *Computer Aided Geometric Design*, vol. 28, no. 8, pp. 449 – 462, 2011.
[98] G. Papandreou and A. Yuille, "Perturb-and-map random fields: Using discrete optimization to learn and sample from energy models," in *Proc. IEEE Int. Conf. on Computer Vision (ICCV)*, Barcelona, Spain, Nov. 2011, pp. 193–200.
[99] T. Tutenel, R. Bidarra, R. M. Smelik, and K. J. D. Kraker, "The Role of Semantics in Games and Simulations," *Comput. Entertain.*, vol. 6, no. 4, pp. 57:1–57:35, Dec. 2008. [Online]. Available: http://doi.acm.org/10.1145/1461999.1462009
[100] D. Kapur and J. L. Mundy, "Geometric Reasoning and Artificial Intelligence: Introduction to the Special Volume," in *Geometric Reasoning*, 1989, pp. 1–14.
[101] J. L. Mundy, O. Faugeras, T. Kanade, C. d'Souza, and M. Sabin, "Object Recognition Based on Geometry: Progress over Three Decades," *Philosophical Transactions: Mathematical, Physical and Engineering Sciences*, vol. 356, pp. 1213–1231, 1998.
[102] C. Brenner, "Constraints for Modelling Complex Objects," in *Int. Archives for Photogrammetry and Remote Sensing*, vol. XXXVI, 3/W24, 2005.
[103] J. Meidow, H. Hammer, M. Pohl, and D. Bulatov, "Enhancement of Generic Building Models by Recognition and Enforcement of Geometric Constraints," in *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, vol. III-3, 2016.
[104] M. Marszalek and C. Schmid, "Semantic Hierarchies for Visual Object Recognition," in *Proc. of Conf. on Computer Vision and Pattern Recognition*, 2007.
[105] G. Griffin and P. Perona, "Learning and using taxonomies for fast visual categorization," in *Proc. of Conf. on Computer Vision and Pattern Recognition*, 2008.
[106] A. Zweig and D. Weinshall, "Exploiting Object Hierarchy:

Combining Models from Different Category Levels," in *Proc. of Int. Conf. on Computer Vision*, 2007.

[107] G. Kramer, G. Bouma, D. Hendriksen, and M. Homminga, "Classifying Image Galleries into a Taxonomy Using Metadata and Wikipedia," in *Proceedings of the 17th International Conference on Applications of Natural Language Processing and Information Systems*, ser. NLDB'12. Berlin, Heidelberg: Springer-Verlag, 2012, pp. 191–196. [Online]. Available: http://dx.doi.org/10.1007/978-3-642-31178-9_20

[108] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A Large-Scale Hierarchical Image Database," in *Proc. of Conf. on Computer Vision and Pattern Recognition*, 2009.

[109] B. Gutmann, A. Kimmig, K. Kersting, and L. d. Raedt, "Parameter learning on probabilistic databases – A least squares approach," in *European Conference on Machine Learning*, 2008.

[110] I. Tsochantaridis, T. Hofmann, T. Joachims, and Y. Altun, "Support Vector Machine Learning for Interdependent and Structured Output Spaces," in *Proceedings of the Twenty-first International Conference on Machine Learning*, ser. ICML '04. New York, NY, USA: ACM, 2004, pp. 104–111. [Online]. Available: http://doi.acm.org/10.1145/1015330.1015341

[111] B. Neumann, A. C. Cohn, D. C. Hogg, and R. Möller, "08091 Abstracts Collection – Logic and Probability for Scene Interpretation," in *Logic and Probability for Scene Interpretation*, ser. Dagstuhl Seminar Proceedings, A. G. Cohn, D. C. Hogg, R. Möller, and B. Neumann, Eds., no. 08091. Dagstuhl, Germany: Schloss Dagstuhl - Leibniz-Zentrum fuer Informatik, Germany, 2008. [Online]. Available: http://drops.dagstuhl.de/opus/volltexte/2008/1648

[112] P. Domingos and W. A. Webb, "A tractable first-order probabilistic logic," in *Proceedings of the Twenty-Sixth AAAI Conference on Artificial Intelligence*, ser. AAAI'12. AAAI Press, 2012, pp. 1902–1909. [Online]. Available: http://dl.acm.org/citation.cfm?id=2900929.2900997

[113] H. Poon and P. Domingos, "Sound and Efficient Inference with Probabilistic and Deterministic Dependencies," in *Proceedings of the 21st National Conference on Artificial Intelligence - Volume 1*, ser. AAAI'06. AAAI Press, 2006, pp. 458–463. [Online]. Available: http://dl.acm.org/citation.cfm?id=1597538.1597612

[114] I. Beltagy and R. J. Mooney, "Efficient Markov Logic Inference for Natural Language Semantics," in *Workshop at the Twenty-Eighth AAAI Conference on Artificial Intelligence*, 2014.

[115] Y. Zhu, C. Zhang, C. Ré, and L. Fei-Fei, "Building a Large-scale Multimodal Knowledge Base for Visual Question Answering," *CoRR*, vol. abs/1507.05670, 2015. [Online]. Available: http://arxiv.org/abs/1507.05670

[116] R. Roscher, "Sequential Learning using Incremental Import Vector Machines for Semantic Segmentation," Ph.D. dissertation, Univerity Bonn, http://hss.ulb.uni-bonn.de/2012/3009/3009.htm, 2012. [Online]. Available: http://hss.ulb.uni-bonn.de/2012/3009/3009.htm

[117] R. Roscher, W. Förstner, and B. Waske, "I2VM: Incremental import vector machines," *Image and Vision Computing*, vol. 30, no. 4-5, pp. 263–278, 2012.

[118] J. Zhu and T. Hastie, "Kernel Logistic Regression and the Import Vector Machine," in *Journal of Computational and Graphical Statistics*. MIT Press, 2001, pp. 1081–1088.

[119] J. C. Platt, "Probabilistic Outputs for Support Vector Machines and Comparisons to Regularized Likelihood Methods," in *Advances in Large Margin Classifiers*. MIT Press, 1999, pp. 61–74.

[120] A. Kendall, V. Badrinarayanan, and R. Cipolla, "Bayesian segnet: Model uncertainty in deep convolutional encoder-decoder architectures for scene understanding," *CoRR*, vol. abs/1511.02680, 2015. [Online]. Available: http://arxiv.org/abs/1511.02680

[121] M. Kampffmeyer, A.-B. Salberg, and R. Jenssen, "Semantic Segmentation of Small Objects and Modeling of Uncertainty in Urban Remote Sensing Images Using Deep Convolutional Neural Networks," in *CVPR Workshop 'Visual Analysis of Satellite to Street Imagery Workshop'*, 2016.

[122] V. Roth and B. Fischer, "Improved functional prediction of proteins by learning kernel combinations in multilabel settings," *BMC Bioinformatics*, vol. 8, pp. S12 – S12, 2007.

[123] M. S. Sorower, "A Literature Survey on Algorithms for Multi-label Learning," Corvallis, OR, Oregon State University, Tech. Rep., 2010.

[124] L. A. Zadeh, "Fuzzy Sets," *Information and Control*, vol. 8, pp. 338–353, 1965.

[125] ——, "The Concept of a Linguistic Variable and its Application to Approximate Reasoning - I," *Information Sciences*, vol. 8, pp. 199–249, 1975.

[126] D. Dubois and H. Prade, "Fuzzy sets and probability: misunderstandings, bridges and gaps," in *International Conference on Fuzzy Systems*, vol. 2, 1993, pp. 1059–1068.

[127] L. A. Zadeh, "Probability Measures of Fuzzy Events," *J. of Math. Analysis and Applications*, vol. 23, no. 2, pp. 421–427, 1968.

[128] M. Navara, "Probability theory of fuzzy events," in *European Society for Fuzzy Logic and Technology*, 2005.

[129] D. Martin, C. Fowlkes, D. Tal, and J. Malik, "A Database of Human Segmented Natural Images and its Application to Evaluating Segmentation Algorithms and Measuring Ecological Statistics," in *Proc. 8th Int'l Conf. Computer Vision*, vol. 2, July 2001, pp. 416–423.

[130] M. Everingham, S. M. A. Eslami, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman, "The Pascal Visual Object Classes Challenge: A Retrospective," *International Journal of Computer Vision*, vol. 111, no. 1, pp. 98–136, Jan. 2015.

[131] J. Fu, S. Levine, and P. Abbeel, "One-Shot Learning of Manipulation Skills with Online Dynamics Adaptation and Neural Network Priors," *CoRR*, vol. abs/1509.06841, 2015. [Online]. Available: http://arxiv.org/abs/1509.06841

[132] J. Bruna and S. Mallat, "Invariant Scattering Convolution Networks," *CoRR*, vol. abs/1203.1513, 2012. [Online]. Available: http://arxiv.org/abs/1203.1513

[133] S. Mallat, "Understanding deep convolutional networks," *Philosophical Transactions of the Royal Society of London A: Mathematical, Physical and Engineering Sciences*, vol. 374, no. 2065, 2016. [Online]. Available: http://rsta.royalsocietypublishing.org/content/374/2065/20150203

[134] K.-Y. Chang, C.-F. Lin, C.-S. Chen, and Y.-P. Hung, "Applying scattering operators for face recognition: A comparative study," in *Pattern Recognition (ICPR), 21st International Conference on*, 2012.

[135] X. Chen, X. Cheng, and S. Mallat, "Unsupervised Deep Haar Scattering on Graphs," in *Advances in Neural Information Processing Systems 27*, Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence, and K. Weinberger, Eds. Curran Associates, Inc., 2014, pp. 1709–1717. [Online]. Available: http://papers.nips.cc/paper/5545-unsupervised-deep-haar-scattering-on-graphs.pdf

[136] G. Tolias, A. Bursuc, T. Furon, and H. Jégou, "Rotation and translation covariant match kernels for image retrieval," *Computer Vision and Image Understanding*, Jun. 2015. [Online]. Available: https://hal.archives-ouvertes.fr/hal-01168525