Efficient Structured Prediction for 3D Indoor Scene Understanding

Alexander G. Schwing
ETH Zurich
aschwing@inf.ethz.ch

Tamir Hazan
TTI Chicago
tamir@ttic.edu

Marc Pollefeys
ETH Zurich
pomarc@inf.ethz.ch

Raquel Urtasun
TTI Chicago
rurtasun@ttic.edu

Abstract

Existing approaches to indoor scene understanding formulate the problem as a structured prediction task focusing on estimating the 3D bounding box which best describes the scene layout. Unfortunately, these approaches utilize high order potentials which are computationally intractable and rely on ad-hoc approximations for both learning and inference. In this paper we show that the potentials commonly used in the literature can be decomposed into pairwise potentials by extending the concept of integral images to geometry. As a consequence no heuristic reduction of the search space is required. In practice, this results in large improvements in performance over the state-of-the-art, while being orders of magnitude faster.

1. Introduction

Recovering the spatial layout of indoor scenes from a single image is an important problem in applications such as personal robotics. Existing approaches typically rely on the *Manhattan world* assumption, which states that there exist three dominant vanishing points (vp) which are orthogonal. They typically formulate the problem as a structured prediction task, which estimates the 3D box that best approximates the scene layout [8, 15, 26]. An example illustrating this is shown in Fig. 1.

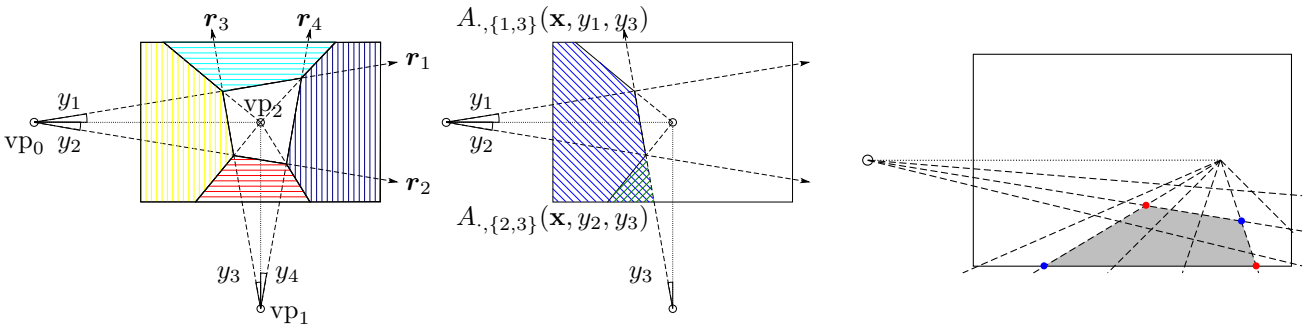
Two different parameterizations have been proposed for this problem, both assuming that the three dominant vanishing points can be reliably detected. In [8, 15], candidate 3D boxes are generated, and inference is formulated in terms of a single high dimensional discrete random variable. Hence, one state of such a variable denotes one candidate 3D layout. This limits significantly the amount of candidate boxes, *e.g.*, only ≈ 1000 candidates are employed in [8]. Contrasting this formulation, Wang *et al*. [26] parameterize the layout with four discrete random variables, that relate to the four degrees of freedom of the problem. In this paper, we adopt the latter parameterization, and model the problem in terms of four random variables that correspond to the angles encoding the rays that originate from the respective vanishing points. As illustrated in Fig. 2(a), these rays fully describe the 3D cuboid, defining the layout.

(a) (b)

Figure 1. Inference result with 3.61% error (colored red), and best discretized solution (labeled blue) is illustrated in (a), while (b) shows a newly synthesized view.

Existing approaches employ potentials based on different image information. Geometric context [11], orientation maps [16] as well as lines in accordance with vanishing points [26] are amongst the most successful cues. While in the single random variable parameterization [8, 15], the complexity is determined directly by the number of candidate boxes, in the parameterization of [26] the complexity is determined by the order of the potentials - the number of variables involved and their size - that encode the image features. These potentials are typically unary, pairwise as well as higher-order (*i.e.*, order four). The order is even higher when reasoning about clutter in the form of hidden variables [26] (*i.e.*, order five) or objects present in the scene that restrict the hypothesis space [15]. While the aforementioned approaches perform well in practice, to tractably handle learning and inference with both parameterizations, reductions on the search space were proposed and/or a limited amount of labelings was considered.

In contrast, in this paper we propose a novel and efficient approach to discriminatively predict the 3D layout of indoor scenes. In particular, we generalize the concept of integral images to "integral geometry," by constructing accumulators in accordance with the vanishing points. We show that utilizing this concept, the potentials and the loss functions frequently used in the literature decompose from order four to order two (*i.e.*, pairwise). As a result, learning and inference is possible without further reduction of the search space. For learning, we exploit the family of structured prediction problems of [7], which encompass structured support vector machines (structured SVMs) and conditional random fields (CRFs) as special cases. Inference is



(a) Parameterization of the problem. (b) Integral geometry for the left-wall. (c) Using accumulators for faces.

Figure 2. Our problem formulation in terms of four random variables y_i is illustrated in (a). Decomposing the third order potential for $\alpha =$ *left-wall* into two shaded second order potentials is shown in (b). A schematic on how "integral geometry" uses accumulators A is given in (c).

done via message-passing.

We evaluate our approach on the main two benchmarks [8, 9] that exist for this task. As shown in our experiments our approach results in significantly better prediction than the state-of-the-art, while being orders of magnitude faster. We are able to perform learning using 50^4 possible labelings in only a few minutes. Moreover, given the potentials, inference over the same hypothesis space (i.e., 50^4 labels) takes 0.15 seconds on average. Although we demonstrate our approach in the problem of predicting the layout of indoor scenes, our integral geometry decomposition is general, and can be applied to other geometric problems such as 3D outdoor scene understanding.

2. Related Work

Over the past few years many approaches have been developed to tackle the problem of semantic scene understanding. The mathematical tools as well as the image features employed by these approaches vary in terms of the particular problem they address. In this section we provide a brief description of the methods employed in the literature.

Most approaches to semantic scene understanding from a single image in outdoor scenarios are qualitative, producing rough 3D in the form of pop-ups [12, 19] as well as 3D image parses that represent the world in terms of blocks [5]. Tretyak *et al*. [23] model the scene as a composition of geometric primitives spanning different layers from low level (edges) over mid-level (line segments, lines and vanishing points) to high level (zenith and horizon). When multiple images in the form of monocular and stereo video sequences are available, quantitative parsings of the road layout and the visible 3D dynamical objects can be constructed by employing generative models of the scene and making use of static and dynamic information [3, 4]. In [1] uncalibrated images where employed to detect objects and to recover the geometry of the scene.

The indoor scenario is more constraint and existing approaches typically rely on vanishing point detection and the Manhattan world properties of man-made indoor scenes. One of the most popular problems in the indoor setting is prediction of the room layout given a single image. The layout is commonly represented in terms of the spatial configuration of the faces of a rectangular 3D cuboid, (*i.e*., left, front and right wall as well as floor and ceiling). This problem is complex, as typical scenes contain objects that partly occlude the walls. The first approach to this problem was developed by [27], where the task is addressed via grouping, *i.e*., edges are grouped into lines, quadrilaterals, and finally depth-ordered planes.

Most recent approaches [8, 15, 26], however, model the problem as inference in a conditional random field (CRF) and learn the parameters using structured SVMs [24]. However, they either only consider a small set of candidate layout boxes, or employ high order potentials which require a very coarse discretization of the space to be computationally tractable. In contrast, in this paper we show that the potentials frequently employed in the literature can be decomposed into sums of pairwise potentials by utilizing the new concept of integral geometry. As a consequence learning and inference is orders of magnitude faster and higher number of states are possible, resulting in more accurate prediction.

Del Pero *et al*. [18] proposed a generative model for the layout. As this model is fairly high dimensional and does not exploit recently developed discriminative image features [11, 16], it results in poor performance when compared to structured prediction approaches.

In [2, 16] a model dealing with more general layouts is considered. [2] addresses scene understanding in the context of a moving camera by combining geometric and photometric cues, *i.e*., stereo photo-consistency, depth cues from structure-from-motion and monocular features. [17] develops order-preserving moves that outperform α-expansion when encoding prior knowledge about the problem, *e.g*., the ceiling should be above the floor. Similar to [10], they train a support vector machine using a large set of features (statistics on location, color, geometry, texture and edges) and demonstrate their optimization technique on the indoor labeling problem. This was fur-

ther extended in [21]. The concept of integral geometry we develop here can be used in this more general setting. However, in this paper we focus on the more standard setting of estimating the room layout as a 3D cuboid.

3. Integral Geometry for Structured Prediction

In this paper we tackle the problem of predicting the room layout of indoor scenes from a single image, and formulate it in terms of the spatial configuration of the faces of a rectangular 3D cuboid, (*i.e.*, left, front and right wall as well as floor and ceiling). In particular, we employ the parameterization of [26] in terms of four random variables.

More formally, let $\mathbf{y}$ be the parameterization of the scene layout. We are interested in learning a linear predictor $\mathbf{w}$, such that, given a new image $\mathbf{x}$, it accurately estimates the room layout by solving the following inference task

$$\text{(inference)} \quad \hat{\mathbf{y}} = \underset{\mathbf{y}}{\operatorname{argmax}}\, \mathbf{w}^T \phi(\mathbf{x}, \mathbf{y}) \tag{1}$$

with $\phi(\mathbf{x}, \mathbf{y})$ denoting a multidimensional feature vector which depends on the image $\mathbf{x}$ and the random variable(s) $\mathbf{y}$. The elements ϕ_r of the feature vector are generally expressed as a sum of non-separable functions, each depending on an arbitrary subset of random variables in $\mathbf{y}$, *i.e.*, potentials involving those variables. The cardinality of those variable subsets for each potential is commonly referred to as its dimensionality of the domain or its order. Consequently, the relations expressed by the features can be visualized as a graphical model. The model structure illustrates that it is generally not possible to accurately predict the state of each random variable independently (without considering the states of other, neighboring random variables). We thus refer to the above inference task as structured prediction.

During learning, the goal is to find weights $\mathbf{w}$ that (i) predict $\mathbf{y}$ as accurately as possible and (ii) generalize well to unseen $\mathbf{x}$. In recent years, many research efforts have been devoted to learning in the structured prediction setting. Two notable frameworks are structured SVMs [22, 24] and CRFs [13]. Existing methods to estimate the room layout [8, 15, 26] typically rely on one of these two approaches.

It remains to answer how to construct the potentials $\phi(\mathbf{x}, \mathbf{y})$ for accurate and efficient prediction. The complexity of the structured prediction problem depends on the order of the potentials involved, and its size, *i.e.*, the number of states. Without objects, the methods presented in [8, 15] have only a single unary potential with inevitably many states. When not considering the clutter denoted by hidden variables, [26] employ potentials of order up to four.

Considering Fig. 2(a), a reasonably dense grid of possible intersections of rays $\boldsymbol{r}_i$,$\boldsymbol{r}_j$, $i \in \{1, 2\}$, $j \in \{3, 4\}$ requires about $N = 50$ angles, *i.e.*, states for each discrete random variable. With the cardinality of the variables being 50, the size of those forth order potentials amounts to 50^4. Similarly, for [8, 15], unary potentials of size N^4 should be considered to define the same hypothesis space. Both methods are computationally demanding and do not scale well in N. The problem is even more severe when considering clutter or when reasoning about objects [15, 26]. In [15] all m objects are ideally connected to each other, *i.e.*, the former unary potential of size $n = 50^4$ is now augmented by as many binary variables as object hypotheses are present in a scene. Hence the potential consists of $n \cdot 2^m$ values. To tractably deal with this space, existing approaches use either fewer states [8], and hence suffer from discretization artifacts, or introduce ad-hoc approximations for learning and inference [15, 26].

In the remainder of the section, we first show how the potentials employed in the literature can be decomposed into pairwise potentials by extending the concept of integral images to integral geometry. We then provide details about the algorithm used to learn the predictor $\mathbf{w}$.

3.1. Integral Geometry

Integral images perform partial computations in accumulators such that the generation of image features at different locations and scales can be performed efficiently by a few accesses to these accumulators. As these accumulators have usually the same size as the image, one commonly refers to them as integral images. They were first introduced by Viola and Jones in their seminal work on real-time face detection [25] to compute Haar-like features, and are nowadays widely used, *e.g.*, in object detection approaches [14].

The potentials employed in the literature to address the problem of indoor scene understanding are typically additive. They count for each face in the cuboid (given a particular configuration of the layout) the number of pixels with a certain label or the probability that such a label appears in the face. Thus, potentials natively depend on three variables for the *left-wall*, *right-wall*, *ceiling* and the *floor* and on four variables for the *front-wall*, as this is the number of variables necessary to define each face.

In this paper, we make the following important observation. In the spirit of integral images, we construct 2D accumulators, each counting features (probabilities) in regions of the space defined by two rays originating from two different vanishing points (*e.g.*, the shaded regions in Fig. 2(b)). We then compute the additive potentials by accessing these accumulators. As each accumulator depends only on two rays, the potentials are constructed as a sum of pairwise factors. We call this decomposition *integral geometry*.

Following Lee et al. [15], we employ geometric context (GC) [11] and orientation maps (OMs) [16] as image information from which we construct the potentials $\phi(\mathbf{x}, \mathbf{y})$. Orientation maps associate a label corresponding to each of the five faces of the 3D cuboid

that are potentially visible in an image, *i.e.*, $\mathcal{F} = \{$*left-wall*, *right-wall*, *ceiling*, *floor*, *front-wall*$\}$. Geometric context provides for every pixel the probability of each surface label which includes *objects* in addition to the five labels in $\mathcal{F}$. Therefore we define $\phi(\mathbf{x}, \mathbf{y})$ as

$$\mathbf{w}^T \phi(\mathbf{x}, \mathbf{y}) = \sum_{\alpha \in \mathcal{F}} \mathbf{w}_{o,\alpha}^T \phi_{o,\alpha}(\mathbf{x}, \mathbf{y}_\alpha) + \sum_{\alpha \in \mathcal{F}} \mathbf{w}_{g,\alpha}^T \phi_{g,\alpha}(\mathbf{x}, \mathbf{y}_\alpha).$$

where $\boldsymbol{y} = \{y_i\}_{i=1}^4 \in \mathcal{Y} = \{1, \ldots, N\}^4$ is the parameterization of the layout in terms of a set of angles, and the subscripts o and g denote OM and GC features respectively. The facets $\alpha \in \mathcal{F}$ are defined in terms of three variables (y_i, y_j, y_k) with $i, j, k \in \{1, \ldots, 4\}$, and $i \neq j \neq k \neq i$, with the exception of the front-wall which requires all four variables. Thus in principle these potentials are of order three and four.

Using the concept of integral geometry, we decompose these factors into potentials of order two. This is illustrated in Fig. 2(b) for the case of the left wall, where

$$\begin{aligned} \phi_{\cdot,\alpha}(\mathbf{x}, \mathbf{y}_\alpha) &= \phi_{\cdot,\{1,2,3\}}(\mathbf{x}, y_1, y_2, y_3) = \\ &= A_{\cdot,\{1,3\}}(\mathbf{x}, y_1, y_3) - A_{\cdot,\{2,3\}}(\mathbf{x}, y_2, y_3), \end{aligned}$$

with A the accumulators that count features. Speaking in colors: for a chosen layout, the area highlighted with yellow in Fig. 2(a) is equal to the difference between the blue and green shaded regions in Fig. 2(b). Computation of each region is illustrated in Fig. 2(c). Similar decompositions in terms of two accumulators are possible for all faces, with the exception of the *front-wall* which requires a larger set of accumulators, but still decomposes into sums of pairwise potentials. In particular,

$$\phi_{\cdot,\textit{front-wall}} = \phi(\mathbf{x}) - \phi_{\cdot,\textit{left-wall}} - \phi_{\cdot,\textit{right-wall}} - \phi_{\cdot,\textit{ceiling}} - \phi_{\cdot,\textit{floor}}$$

where each of the $\phi_{\cdot,\alpha}$ is decomposed into potentials of order two. $\phi(\mathbf{x})$ denotes a global constant potential which is independent of the variables. It counts all the features in the image. The vector A measures a quantity within the respective face. For orientation maps, $A_{o,\cdot}$ is a five dimensional vector counting how many pixels are assumed to correspond to each cuboid face. For geometric context, $A_{g,\cdot}$ is a six dimensional vector that counts the probability that each pixel belongs to the different faces as well as objects. The weight vector $\mathbf{w}$ measuring the importance of each feature is hence 55-dimensional. While this is the particular choice of potentials we make in this paper, similar accumulators can be used to decompose the potentials employed in other approaches, *e.g.*, [26].

It is important to note that we compute the accumulators efficiently by sorting every pixel into a bin of an image grid spanned by any combination of two vanishing points. Treating these histograms as $N \times N$ sized pictures, and computing integral images on them, provides a method that allows to obtain the accumulators A with a complexity of $O(M + N^2)$, with M being the number of pixels of the image and N indicating the quantization of the histograms.

3.2. Efficient Structured Prediction

CRFs [13] and structured SVMs [22, 24] are typically employed to solve tasks with complex dependencies between the discrete random variables. Recently, Hazan and Urtasun [7] unified these two methods into a single structured prediction framework using the soft-max function which smoothly approximates the hinge loss function of structured SVMs. Given a dataset D of training pairs $(\mathbf{x}, \mathbf{y})$, learning in this framework is done by solving the following optimization problem

$$\min_{\mathbf{w}} \sum_{(\mathbf{x},\mathbf{y}) \in D} \ln Z_\epsilon(\mathbf{x}, \mathbf{y}) - \mathbf{w}^T \mathbf{d} + \frac{C}{p} ||\mathbf{w}||_p^p \qquad (2)$$

where $\mathbf{d} = \sum_{(\mathbf{x},\mathbf{y}) \in D} \phi(\mathbf{x}, \mathbf{y})$ is the vector of empirical means, C and p are constants, and $\ln Z_\epsilon(\mathbf{x}, \mathbf{y})$ is the one parameter extension of the log-partition function

$$\ln Z_\epsilon(\mathbf{x}, \mathbf{y}) = \epsilon \ln \sum_{\hat{\mathbf{y}} \in \mathcal{Y}} \exp\left(\frac{\ell(\mathbf{y}, \hat{\mathbf{y}}) + \mathbf{w}^T \phi(\mathbf{x}, \hat{\mathbf{y}})}{\epsilon} \right) \qquad (3)$$

with $\ell(\mathbf{y}, \hat{\mathbf{y}})$ the loss of predicting $\hat{\mathbf{y}}$ instead of $\mathbf{y}$.

Interpreting the log-partition function given in Eq. (3) as a $\frac{1}{\epsilon}$-norm, we observe that the optimization problem in Eq. (2) when setting $\epsilon = 0$ is given by

$$\min_{\mathbf{w}} \sum_{(\mathbf{x},\mathbf{y}) \in D} \max_{\hat{\mathbf{y}} \in \mathcal{Y}} \left(\ell(\mathbf{y}, \hat{\mathbf{y}}) + \mathbf{w}^T \phi(\mathbf{x}, \hat{\mathbf{y}}) \right) - \mathbf{w}^T \mathbf{d} + \frac{C}{p} ||\mathbf{w}||_p^p \qquad (4)$$

This formulation is identical to the margin rescaling approach of [24]. Hence we recover the structured SVM for $\epsilon = 0$ and the CRF when $\epsilon = 1$. However, the parameter ϵ is not restricted to the interval $[0, 1]$.

Dealing with the partition function (*i.e.*, $\ln Z_\epsilon(\mathbf{x}, \mathbf{y})$) is hard, as it involves a sum or a max operation over exponentially many labels. We take advantage of the primal-dual message-passing approach of [7] to approximate the problem and solve the approximated problem exactly.

Besides the interpretation of smoothly approximating the non-smooth $\max$ function involved in the structured SVM objective, ϵ can be seen as a parameter that adjusts the ratio between loss and regularization.

We use the per-pixel classification error (*i.e.*, the percentage of pixels that have been wrongly predicted as being part of another face) as our loss $\ell(\mathbf{y}, \hat{\mathbf{y}})$. The concept of integral geometry introduced above allows us to also decompose the loss into at most pairwise potentials. This is important as the complexity of the learning depends on the order of the loss potentials in the same manner as it depends on the order of the image feature potentials.

We explore extensively the influence of parameters C and ϵ in our experiments below. Importantly, the decomposition of the loss and the potentials in conjunction with the structured prediction framework of [7] allow us to perform learning with 50^4 possible labels in only a few minutes.

We now derive our inference algorithm for solving Eq. (1). Exploiting the previously mentioned graphical model structure, each element r of our feature vector $\phi(\mathbf{x}, \mathbf{y}) = [\phi_1(\mathbf{x}, \mathbf{y}), \ldots, \phi_F(\mathbf{x}, \mathbf{y})]^T$ can be written as

$$\phi_r(\mathbf{x}, \mathbf{y}) = \sum_{i \in V_{r,x}} \phi_{r,i}(\mathbf{x}, y_i) + \sum_{\gamma \in E_{r,x}} \phi_{r,\gamma}(\mathbf{x}, \mathbf{y}_\gamma),$$

with F the total number of features ($F = 55$ in our experiments), $V_{r,x}$ and $E_{r,x}$ denoting the nodes and cliques of the graphical model corresponding to feature r of sample $\mathbf{x}$. Relating this general notation to the example of the left wall illustrated in Fig. 2(b), we obtain $E_{r,x} = \{\{1,3\},\{2,3\}\}$ and $V_{r,x} = \{1,2,3\}$ with $\phi_{r,i} = 0 \; \forall i \in V_{r,x}$ and $\phi_{r,\gamma}$ corresponding to entries in the respective accumulators A. When making use of the general graphical model structure, we can rewrite the maximization of $\mathbf{w}^T\phi(\mathbf{x}, \mathbf{y})$ in Eq. (1) as the following program

$$\max_{\mathbf{y}} \sum_{i \in V_x, \beta : i \in V_{\beta,x}} w_\beta \phi_{\beta,i}(\mathbf{x}, y_i) + \sum_{\gamma \in E_x, \beta : \gamma \in E_{\beta,x}} w_\beta \phi_{\beta,\gamma}(\mathbf{x}, \mathbf{y}_\gamma), \tag{5}$$

with $V_x = \bigcup_r V_{r,x}$ and $E_x = \bigcup_r E_{r,x}$ denoting the nodes and cliques of the union hypergraph, subsuming all features of sample $\mathbf{x}$. We employ convex max product belief propagation [6, 20] to optimize this problem. The complexity of the message passing depends on the size of the potentials being $O(N^2)$ when applying the decomposition described previously. This is extremely efficient as inference in our model takes on average 0.15 seconds per image.

4. Experimental Evaluation

We evaluate our approach on the data sets of [8, 9]. The layout data set [8] contains 314 images with ground truth annotation of layout faces. We employed the vanishing point detection of [8], which failed in 9 training images and was successful for all test images, 105 in total. The bedroom data set [9] contains 309 labeled images, split into training and test sets of size 181 and 128 respectively. In accordance with previous work on those data sets, we use a pixel based error measure, counting the percentage of pixel that disagree with the provided ground truth labeling. In the following, we first compare our approach to the state-of-the-art and evaluate the performance on a family of (approximated) structured prediction problems [7, 24]. We then investigate the dependency of the prediction accuracy on the training time as well as on the size of the training set, and provide some experiments regarding a simple object reasoning before concluding by showing success and failure

	OM	GC	OM + GC
[11]	-	28.9	-
[8] (a)	-	26.5	-
[8] (b)	-	21.2	-
[26]	22.2	-	-
[15]	24.7	22.7	18.6
Ours (SVM$^{\text{struct}}$)	**18.7**	**15.4**	**14.0**
Ours (median struct-pred)	**18.9**	**15.6**	**14.0**
Ours (best struct-pred)	**18.6**	**15.4**	**13.6**

Table 1. Comparison to state-of-the-art that uses the same image information on the layout data set of [8]. Pixel classification error is given in %.

	[18]	[11]	[8](a)	Ours (best/median)
w/o box	29.59	23.04	22.94	**16.46/16.93**
w/ box	26.79	-	22.94	**15.19/15.59**

Table 2. Comparison to state-of-the-art on the bedroom data set [9]. Pixel classification error is given in %.

cases. During learning, unless otherwise stated, we use a relative duality gap $\left(\frac{\text{primal}-\text{dual}}{\text{primal}}\right)$ of $1e{-}5$ or at most 500 iterations as stopping criteria for the optimization. All experiments were performed on an 8-core, 2.4GHz Intel Xeon CPU. We initialize the predictor $\mathbf{w}$ to the all ones vector. Note that ideally the parameters C and ϵ should be chosen after cross-validation. As we directly employed the features provided by [8], which are learned on the training set, we use the training/test set split given in [8] and [9]. We note that the median of all the results obtain with all ranges of parameters C and ϵ, has a small standard deviation and compares favorably to the state-of-the-art. This demonstrates that cross-validation (if possible) will also result in state-of-the-art performance.

Comparison to state-of-the-art: We first compare our approach to the state-of-the-art on the layout dataset [8]. Similar to [15], we report results when using different sets of image features, *i.e.*, orientation maps (OM), geometric context (GC), and both (OM+GC). We denote by [8] (a), when the GC features are used to estimate the layout, and by [8] (b), when the layout is used to re-estimate the GC features, and these new features are used to improve the layout. As shown in Tab. 1, our approach is able to significantly outperform the state-of-the-art in all scenarios, with our smallest error rate when using all features being 13.59%. We improve the state-of-the-art by 3.6% for the OM features, by 5.8% for the GC features and by 5.0% when combining both feature cues.

On the bedroom data set [9] we observe a similar effect which is summarized in the row of Tab. 2 denoted by 'w/o box.' As in this data set there are no results that distinguish different image cues, we simply provide results using both cues. Our approach improves state-of-the-art by 6.48%.

CRFs vs. structured SVMs: We next investigate the trade-off between regularization parameter C and the

ϵ \ C	1e2	1	1e-2	1e-4	1e-6
10.0000	30.92	17.53	13.81	14.24	14.46
1.0000	23.95	16.26	14.46	14.86	14.86
0.1000	17.64	13.69	14.83	14.80	14.80
0.0100	15.83	**13.59**	14.20	14.25	14.25
0.0010	15.46	13.82	14.00	13.85	13.85
0.0001	16.04	14.09	13.95	13.96	13.97
0	15.72	13.70	13.82	13.91	13.98

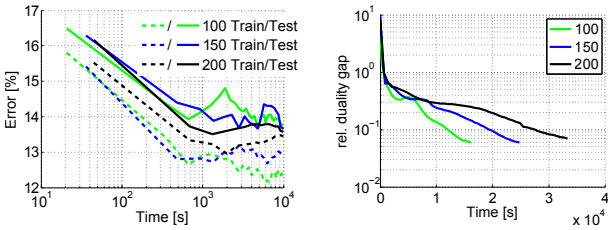
Table 3. Test set percentage pixel error on the layout data set [8] when using both feature types.

ϵ \ C	1e2	1	1e-2	1e-4	1e-6
10.0000	27.28	19.54	17.43	**16.46**	16.49
1.0000	22.66	17.87	16.55	16.83	16.83
0.1000	19.41	17.43	16.74	16.91	16.96
0.0100	17.98	16.48	17.04	16.86	16.86
0.0010	17.92	17.18	16.95	16.83	16.87
0.0001	17.95	17.06	17.07	16.80	16.71
0	18.01	16.97	17.25	17.04	17.01

Table 4. Test set pixel classification error in % with both (OM) features and (GC) on the bedroom data set of [9].

weight of the loss ϵ. Note that $\epsilon = 0$ recovers the structured SVM, and $\epsilon = 1$ results in the CRF. The per pixel classification errors are shown in Tab. 3 for OM+GC features on the layout dataset. For the bedroom data set results are provided in Tab. 4. In our experiments we note an approximately equal performance of CRFs and structured SVMs. Considering ϵ as a weight for the loss, it is not astonishing that neither CRFs nor structured SVMs achieve the best performance. An appropriate ratio between loss and regularization in the primal domain corresponds to a ratio between uncertainty+prior and moment matching in the dual formulation. Investigating the dual therefore reveals that an increasing weight for the moment matching ($C \rightarrow 0$) can reduce generalization performance. Nevertheless, large parts of the $C - \epsilon$ domain show about the same accuracy.

Another learning approach: If we restrict ourselves to $\epsilon = 0$, SVM$^{\text{struct}}$ [24] is a common learning alternative. Hence, we compare our implementation of a structured prediction algorithm to this publicly available framework, where we applied the margin rescaling approach on the layout data set. When using both OM and GC features, [24] results in a pixel wise error of 14.0% as denoted by "Ours (SVM$^{\text{struct}}$)" in Tab. 1. This error is consistent with the results of our implementation for $\epsilon = 0$, given in Tab. 3. As for the results presented in Tab. 3, a relative duality gap of $1e{-}5$ was the provided stopping criterion. The implementation of [24] is not parallelized. To speed up learning, we employ our efficient convex belief propagation implementation with integral geometry in the inner loop.



(a) (b)

Figure 3. The decrease of the error on the training and the test set of [8] for different sizes of the training set is illustrated in (a). The decrease of the relative duality gap is given in (b). The parameters are $\epsilon = 0.01$ and $C = 1$. Both, OM and GC features are used.

	OM	GC	OM + GC
[26]	20.1	-	-
[15]	19.5	20.2	16.2
Ours (SVM$^{\text{struct}}$)	**17.3**	**14.6**	**13.8**
Ours (best struct-pred)	**17.1**	**14.2**	**12.8**

Table 5. For different feature cues we compare to all state-of-the-art using object reasoning on the layout data set [8].

Time for training and inference: We next investigate the training time required to obtain accurate models. Fig. 3 shows the decrease of the test and training error (Fig. 3(a)) as well as the duality gap (Fig. 3(b)) as a function of time for $\epsilon = 0.01$ and $C = 1$. After less than 15 minutes of training with both OM and GC features we obtain test set errors below 14%. Less than five minutes of training are required to outperform the state-of-the-art. If we run inference till convergence, *i.e.*, a duality gap of less than $1e{-}5$ or a maximum of 500 iterations, the average time to estimate the 3D layout of one test set scene given the features is 0.15 seconds. Among the 105 test images in the layout data set, inference on eleven scenes did not converge within 500 iterations.

Training set size: Fig. 3 illustrates the behavior for different training set sizes. As expected, we observe that an increasing training set size reduces the test error. The difference between 100 and 200 training instances is small. We also note that training with less images is of course computationally less expensive and lower relative duality gap rates are obtained faster. We further show in Fig. 3(a) that the difference between training error (dashed line) and test error (solid line) decreases when increasing the set size.

Ground truth assignments: For learning, we need to obtain ground truth states from the provided pixel labels. This is achieved by exhaustively searching all possible state combinations for the best loss $\ell(\mathbf{y}, \hat{\mathbf{y}})$. Alternatively, human labeled ground truth could of course be leveraged. Ground truth assignments being the best possible discretized solution are shown in blue color in Fig. 1,4,5.

Simple object reasoning: Lee et al. [15] show that reasoning about objects in the scene can improve performance

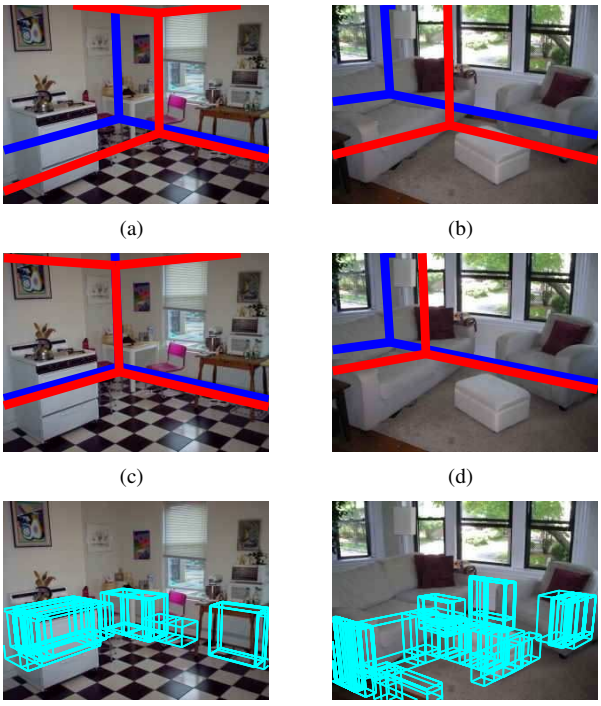


Figure 4. Ground truth and inference results with 16.87% and 23.17% error in (a) and (b). Our simple object reasoning improves to the 7.80% and 10.72% error results shown in (c) and (d). The object hypotheses are illustrated in (e) and (f).

on the layout estimation. Therefore, we use the orientation maps to generate hypotheses of objects which are assumed to be located on the ground plane. We create additional potentials that linearly penalize the angles intersecting with those hypothesized objects. This type of object reasoning adds two unary and four pairwise potentials to the feature vector, which is now 61-dimensional. Tab. 5 compares our simple object reasoning to the more complex approaches of [15, 26] for the layout data set. We obtain our best prediction accuracy of 12.8% which improves the state-of-the-art by 3.4%. For the bedroom data set we provide the best result and the median in the row of Tab. 2 denoted 'w/ box.' Visual improvements are illustrated in Fig. 4.

Qualitative evaluation: We show qualitative results in Fig. 5. For each scene we provide three images. The first one shows our estimation (red) overlaying the original image as well as the best ground truth labels given the detected vanishing points (blue). The next two images show color-coded OM and GC features respectively, with red, green, blue, yellow and cyan indicating floor, front-wall, right-wall, left-wall and ceiling. Fig. 5(k)-(l) depict failure modes. On both data sets we identify two sources of errors, problematic features and wrong vanishing points. A result with a wrong vanishing point is illustrated in Fig. 5(l). Although our prediction is close to ground truth, the error is 42.1%.

5. Conclusions

In this paper we have addressed the problem of recovering the scene layout in the form of a 3D parametric box given a single image. We have introduced the novel concept of *integral geometry* and show that using this concept, the potentials used in the literature can be decomposed into pairwise potentials. This results in an efficient structured prediction framework which allows to solve the problem without any ad-hoc approximations. This is very important in practice, as our approach significantly outperforms the state-of-the-art in both accuracy and time. We plan to investigate more sophisticated object models, which will hopefully allow for navigation on indoor scenarios utilizing only visual sensors.

References

[1] S. Bao and S. Savarese. Semantic Structure from Motion. In *Proc. CVPR*, 2011.

[2] A. Flint, D. Murray, and I. Reid. Manhatten Scene Understanding Using Monocular, Stereo, and 3D Features. In *Proc. ICCV*, 2011.

[3] A. Geiger, M. Lauer, and R. Urtasun. A Generative Model for 3D Urban Scene Understanding from Movable Platforms. In *Proc. CVPR*, 2011.

[4] A. Geiger, C. Wojek, and R. Urtasun. Joint 3D Estimation of Objects and Scene Layout. In *Proc. NIPS*, 2011.

[5] A. Gupta, A. A. Efros, and M. Hebert. Blocks World Revisited: Image Understanding Using Qualitative Geometry and Mechanics. In *Proc. ECCV*, 2010.

[6] T. Hazan and A. Shashua. Norm-Prodcut Belief Propagation: Primal-Dual Message-Passing for LP-Relaxation and Approximate-Inference. *Trans. Information Theory*, 2010.

[7] T. Hazan and R. Urtasun. A Primal-Dual Message-Passing Algorithm for Approximated Large Scale Structured Prediction. In *Proc. NIPS*, 2010.

[8] V. Hedau, D. Hoiem, and D. Forsyth. Recovering the Spatial Layout of Cluttered Rooms. In *Proc. ICCV*, 2009.

[9] V. Hedau, D. Hoiem, and D. Forsyth. Thinking Inside the Box: Using Appearance Models and Context Based on Room Geometry. In *Proc. ECCV*, 2010.

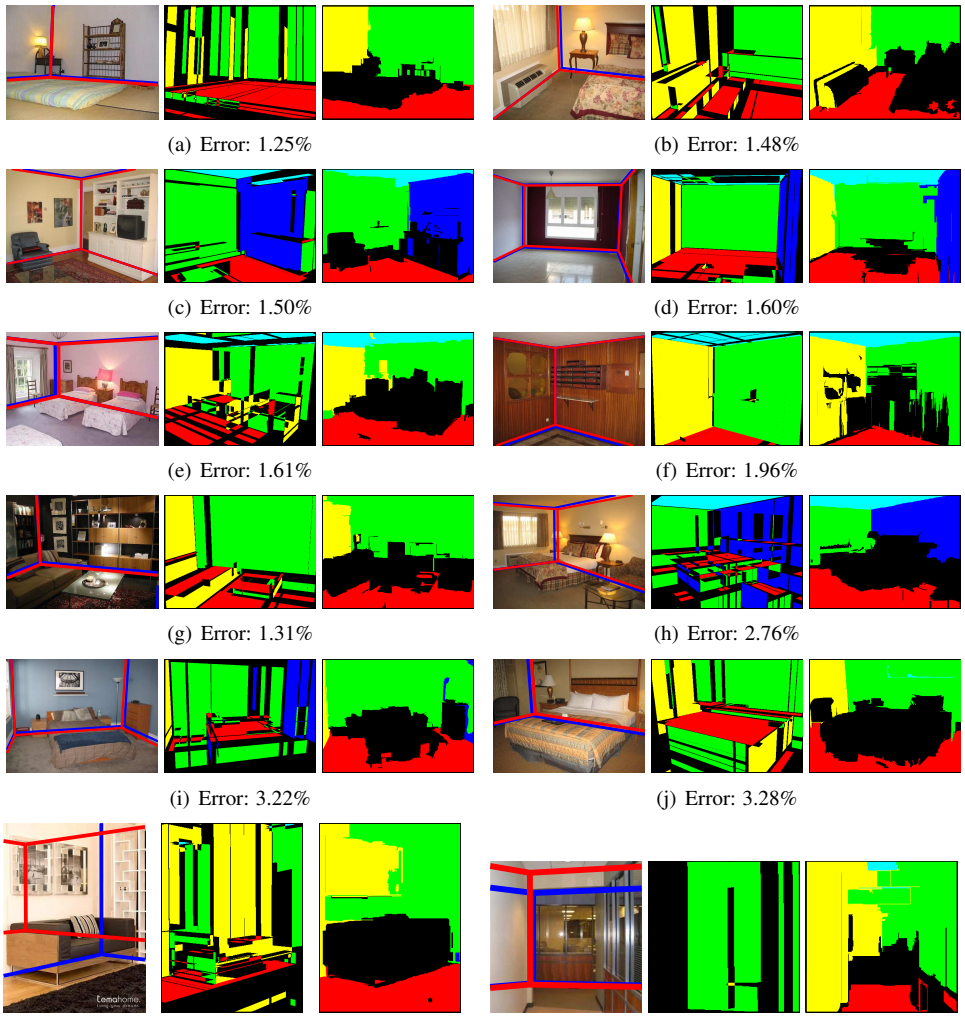
[10] D. Hoiem, A. A. Efros, and M. Hebert. Automatic Photo Pop-up. In *Siggraph*, 2005.

[11] D. Hoiem, A. A. Efros, and M. Hebert. Recovering surface layout from an image. *IJCV*, 2007.

[12] D. Hoiem, A. A. Efros, and M. Hebert. Putting Objects in Perspective. *IJCV*, 2008.

[13] J. Lafferty, A. McCallum, and F. Pereira. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proc. ICML*, 2001.

[14] C. H. Lampert, M. B. Blaschko, and T. Hofmann. Efficient Subwindow Search: A Branch and Bound Framework for Object Localization. *PAMI*, 2009.



(a) Error: 1.25% (b) Error: 1.48%

(c) Error: 1.50% (d) Error: 1.60%

(e) Error: 1.61% (f) Error: 1.96%

(g) Error: 1.31% (h) Error: 2.76%

(i) Error: 3.22% (j) Error: 3.28%

(k) Error: 48.56% (l) Error: 42.11%

Figure 5. (a)-(f): The six best detection results (OM+GC features) on the layout data set with pixel classification errors as indicated. (g)-(j): The four best detection results (OM+GC features) on the bedroom data set with pixel classification errors as indicated. (k)-(l): Two failing detections on the layout data set (OM+GC features) with pixel classification errors as indicated.

[15] D. C. Lee, A. Gupta, M. Hebert, and T. Kanade. Estimating Spatial Layout of Rooms using Volumetric Reasoning about Objects and Surfaces. In *Proc. NIPS*, 2010.

[16] D. C. Lee, M. Hebert, and T. Kanade. Geometric Reasoning for Single Image Structure Recovery. In *Proc. CVPR*, 2009.

[17] X. Liu, O. Veksler, and J. Samarabandu. Graph Cut with Ordering Constraints on Labels and its Applications. In *Proc. CVPR*, 2008.

[18] L. Pero, J. Guan, E. Brau, J. Schlecht, and K. Barnard. Sampling Bedrooms. In *Proc. CVPR*, 2011.

[19] A. Saxena, S. Chung, and A. Y. Ng. 3-D Depth Reconstruction from a Single Still Image. *IJCV*, 2008.

[20] A. G. Schwing, T. Hazan, M. Pollefeys, and R. Urtasun. Distributed Message Passing for Large Scale Graphical Models. In *Proc. CVPR*, 2011.

[21] E. Strekalovskiy and D. Cremers. Generalized Ordering Constraints for Multilabel Optimization. In *Proc. ICCV*, 2011.

[22] B. Taskar, S. Lacoste-Julien, and M. I. Jordan. Structured prediction, dual extragradient and Bregman projections. *JMLR*, 2006.

[23] E. Tretyak, O. Barinova, P. Kohli, and V. Lempitsky. Geometric Image Parsing in Man-Made Environments. *IJCV*, 2012.

[24] I. Tsochantaridis, T. Joachims, T. Hofmann, and Y. Altun. Large margin methods for structured and interdependent output variables. *JMLR*, 2005.

[25] P. Viola and M. Jones. Robust real-time face detection. *IJCV*, 2004.

[26] H. Wang, S. Gould, and D. Koller. Discriminative Learning with Latent Variables for Cluttered Indoor Scene Understanding. In *Proc. ECCV*, 2010.

[27] S. Yu, H. Zhang, and J. Malik. Inferring Spatial Layout from A Single Image via Depth-Ordered Grouping. In *Proc. Workshop on Perceptual Organization in Computer Vision*, 2008.