Efficient Exact Inference for 3D Indoor Scene Understanding

Alexander G. Schwing[1] and Raquel Urtasun[2]

[1] ETH Zurich
[2] TTI Chicago

Abstract. In this paper we propose the first exact solution to the problem of estimating the 3D room layout from a single image. This problem is typically formulated as inference in a Markov random field, where potentials count image features (*e.g.*, geometric context, orientation maps, lines in accordance with vanishing points) in each face of the layout. We present a novel branch and bound approach which splits the label space in terms of candidate sets of 3D layouts, and efficiently bounds the potentials in these sets by restricting the contribution of each individual face. We employ integral geometry in order to evaluate these bounds in constant time, and as a consequence, we not only obtain the exact solution, but also in less time than approximate inference tools such as message-passing. We demonstrate the effectiveness of our approach in two benchmarks and show that our bounds are tight, and only a few evaluations are necessary.

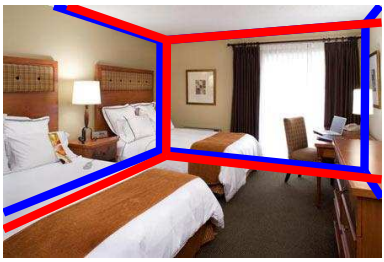
1 Introduction

3D scene understanding is an important component in applications such as autonomous driving and personal robotics. Existing approaches that perform inference from a single image can be categorized into those that attempt to estimate the layout of outdoor scenes and those that tackle the indoor setting. In outdoor scenarios, existing approaches try to infer 3D information in the form of photo pop-ups [1, 2], qualitative parsings of the scene under geometric constraints [3, 4] or reason about the layout of road intersections [5].

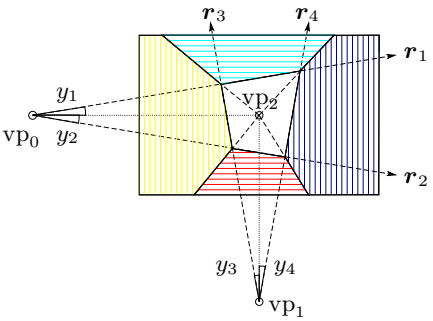
The indoor scenario is more constrained, as the *Manhattan world* assumption typically holds. This assumption states that there exist three dominant vanishing points which are orthogonal. The problem of inferring the 3D layout of indoor scenes is commonly formulated as a structured prediction task, which estimates the 3D box that best approximates the scene layout [6–9]. A generalization of this setting from rooms to corridors with arbitrary number of frontal walls has also been explored [10, 11].

In this paper, we focus on the more popular setting of estimating the layout of rooms from single images as illustrated in Fig. 1(a). The two prevalent parameterizations targeting this problem assume that the three dominant vanishing points (vp) are reliably detected. In [6, 7], a single high dimensional discrete

A. Fitzgibbon et al. (Eds.): ECCV 2012, Part VI, LNCS 7577, pp. 299–313, 2012.



(a) 3D layout prediction.



(b) Parameterization of the problem.

Fig. 1. In (a) we illustrate the task of predicting the room layout given a single image with the blue overlay illustrating the best possible result and red indicating an obtained prediction of 6.47% error. The parameterization of the problem is detailed in (b).

random variable with each state denoting a different candidate 3D box is employed. However, only a few candidate layouts were considered as this setting does not allow to decompose the problem. Contrasting this univariate formulation, Wang *et al*. [8] propose a parameterization of the room layout based on four discrete random variables related to the four degrees of freedom of the underlying prediction problem. As shown in Fig. 1(b), these four variables correspond to rays which fully describe the 3D cuboid that defines the layout.

Nearly all existing methods make use of different sources for image information. Geometric context [12], orientation maps [11] and lines corresponding to a particular vanishing point [8] are most successfully employed as image cues. While the complexity is determined directly by the number of candidate layouts when working with the univariate parameterization [6, 7], we obtain a complexity dependence on the dimensionality of the domain of the potentials representing the image features, *i.e*., the order of the potentials - the number of variables involved and their size. In recent work [9], the concept of integral geometry was introduced, and it was shown that using this concept the potentials employed in the literature are decomposable into sums of pairwise potentials. This results in orders of magnitude faster inference than previously published layout estimation approaches.

In this paper we go a step forward in this direction and show that not only the exact solution to this problem can be obtained, but also in less time than the efficient but non-exact message-passing of [9]. In particular, we derive a novel branch and bound approach to this problem which splits the label space in terms of candidate sets of 3D layouts, and bounds the potentials in these sets by restricting the contribution of each individual face. By employing integral geometry, we are able to compute the bounds in constant time, yielding the exact solution to the problem in a fraction of a second on a single core computer. We demonstrate the effectiveness of our approach using the layout dataset of [6] as well as the bedroom dataset of [13], and show that our bounds are tight, and only a few evaluations are necessary.

2 Related Work

Most outdoor semantic scene understanding approaches produce either only qualitative results [3] or a mild level of understanding in the form of semantic labels [14, 15], object detections [16] or rough 3D [17, 18]. An exception to this is [5] which relies on short video sequences, or [19] which relies on uncalibrated image pairs. While outdoor scenarios remain fairly unexplored in computer vision due to their difficulty, estimating the 3D layout of indoor scenes has experienced increased popularity in the past few years. This can be mainly attributed to the success of novel structured prediction methods as well as the fact that indoor scenes behave mostly as 'Manhattan worlds,' *i.e.*, edges on the image can be associated with parallel lines defined in terms of the three dominant vanishing points which are orthogonal. The Manhattan world assumption has frequently been used in the past for tasks such as orientation estimation [20] and regularity detection [21].

Indoor scene understanding approaches reason about the layout of rooms in the form of 3D cuboids [6–8, 13, 22]. In [6, 7], the problem is treated as the one of selecting a cuboid from a set of candidate 3D layouts. This limits the performance as only a small number of layouts is considered. Wang *et al.* [8] parameterize the problem with four random variables, which represent the angles whose rays originate from two different vanishing points and thus define the faces of the 3D cuboid. While more effective, they rely on higher order potentials with up to four involved variables. As they additionally reason about clutter, the dimensionality of these potentials is in fact even higher. Therefore, in [8], they make use of an iterated conditional modes (ICM) algorithm to tractably deal with the complex potentials. However, this algorithm is neither efficient nor globally convergent as ICM can get stuck in local optima.

Structure prediction approaches employ potentials based on different image information. Geometric context [12], orientation maps [11] as well as lines in accordance with vanishing points [8] are amongst the most successful cues. The potentials employed in the literature count these features for each facet of the cuboid, *i.e.*, left wall, right wall, ceiling, floor, front wall. Schwing *et al.* [9] recently showed that these potentials are decomposable into sums of pairwise potentials by using the concept of integral geometry. This concept is analogous to integral images, but the accumulators are formed in terms of pairs of rays that describe the cuboid. As a result, inference in this model is orders of magnitude faster than previous approaches. The inference algorithm used to find the most likely configuration is however approximate. Therefore, we would like to know how far we are from estimating the true maximum a-posteriori (MAP) solution of the problem. By leveraging branch and bound techniques we show in this paper how to obtain the exact solution of the problem in less time. Interestingly, we find that [9] obtained the optimum for almost all cases.

Del Pero *et al.* [23, 24] proposed to solve the layout prediction task using a generative method. The resulting performance is poor when compared to structured prediction approaches as their model is fairly complex and does not exploit the recently developed discriminative image features of [11, 12].

Algorithm 1. branch and bound (BB) inference

put pair $(\bar{f}(\mathcal{Y}), \mathcal{Y})$ into queue and set $\hat{\mathcal{Y}} = \mathcal{Y}$
repeat
  split $\hat{\mathcal{Y}} = \hat{\mathcal{Y}}_1 \times \hat{\mathcal{Y}}_2$ with $\hat{\mathcal{Y}}_1 \cap \hat{\mathcal{Y}}_2 = \emptyset$
  put pair $(\bar{f}(\hat{\mathcal{Y}}_1), \hat{\mathcal{Y}}_1)$ into queue
  put pair $(\bar{f}(\hat{\mathcal{Y}}_2), \hat{\mathcal{Y}}_2)$ into queue
  retrieve $\hat{\mathcal{Y}}$ having highest score
until $|\hat{\mathcal{Y}}| = 1$

More recently, [10, 11] have tackled the problem of inferring the 3D layout of corridors. While this is arguably simpler due to the fact that the level of clutter in corridors is much lower, it requires doing inference over a larger set of random variables. Existing approaches typically reason at the pixel level, having thousands of variables. In contrast, in this paper we tackle the more popular room layout estimation problem.

Branch and bound techniques have been used in the past to tackle computer vision problems. The seminal work of [25] shows that for a family of object detectors (*e.g.*, bag of words (BoW) with linear kernels, spatial pyramids, intersection kernels), one can compute the MAP solution without relying on approximate sliding window approaches. Those methods are approximate as only a subset of the possible bounding boxes are typically considered (*e.g.*, fixed aspect ratio). Branch and bound was also employed in [26] to bound the deformation cost of deformable part-based models [16]. In contrast, in this paper we derive bounds for the problem of 3D indoor scene understanding.

3 Exact Layout Inference

We tackle the problem of predicting the room layout of indoor scenes from 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. In order to simplify the inference process, following existing approaches [6–9] we rely on vanishing point detection and the Manhattan world properties of man-made indoor scenes.

We define a 3D room layout via four rays r_i originating from two distinct vanishing points. Considering the geometry depicted in Fig. 1(b) we can construct a hypothesis layout from two 2D points given by the intersections of r_1 with r_3 and r_2 with r_4. Let the tupel $y = (y_1, \ldots, y_4) \in \mathcal{Y} = \prod_{i=1}^{4} \{[y_{i,min}, y_{i,max}]\}$ be the parameterization of the scene layout in terms of the angles that form these four rays [8]. Every angle y_i with $i \in \{1, \ldots, 4\}$ lies within the interval denoted $[y_{i,min}, y_{i,max}]$ and $\mathcal{Y}$ indicates the product space of all four angles. To obtain an accurate prediction $\hat{y}$ given an image x we need to solve the following inference task:

$$\hat{y} = \operatorname*{argmax}_{y \in \mathcal{Y}} \mathbf{w}^T \phi(x, y). \tag{1}$$

The model parameters $\mathbf{w}$ commonly referred to as weights can be obtained with structured prediction learning algorithms such as structured SVMs [27, 28], conditional random fields [29] or approximate structured prediction [30].

Following Lee *et al.* [7], we employ geometric context (GC) [6] and orientation maps (OMs) [11] as image information from which we construct the feature vector $\phi(x, y)$. For a subset of the pixels, orientation maps provide a label corresponding to one of the five faces of the 3D cuboid that is potentially visible in the image, *i.e.*, $\mathcal{F} = \{\textit{left-wall}, \textit{right-wall}, \textit{ceiling}, \textit{floor}, \textit{front-wall}\}$. Geometric context [6], on the other hand, provides for every pixel the probability that this pixel belongs to each surface label including *objects* in addition to the five labels in $\mathcal{F}$. Given these image features, potentials are constructed by counting for each face each feature type. We thus define $\phi(x, y)$ as a sum of potentials

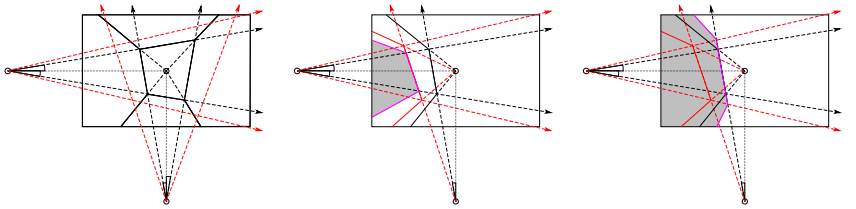
$$\mathbf{w}^T \phi(x, y) = \sum_{\alpha \in \mathcal{F}} \mathbf{w}_{o,\alpha}^T \phi_{o,\alpha}(x, y_\alpha) + \sum_{\alpha \in \mathcal{F}} \mathbf{w}_{g,\alpha}^T \phi_{g,\alpha}(x, y_\alpha), \tag{2}$$

where the subscripts o and g denote OM and GC features respectively. Note that the vectors $\mathbf{w}_{o,\alpha}$ and $\mathbf{w}_{g,\alpha}$ consist of 5 elements $w_{o,\alpha,i}$ with $i \in \{1, \ldots, 5\}$ for each face $\alpha \in \mathcal{F}$ in the case of orientation maps and 6 elements $w_{g,\alpha,i}$ with $i \in \{1, \ldots, 6\}$ per face $\alpha \in \mathcal{F}$ for geometric context.

We now describe how branch and bound is employed for our problem. We start with a trivial set (*i.e.*, all possible layouts $\mathcal{Y}$), and at any given branch and bound iteration we have a priority queue where the considered sets are ordered in terms of a quality bound function which upper bounds the maximum score that any layout member of that set can possibly achieve. The best candidate $\hat{\mathcal{Y}}$ of the layout sets within the queue is considered. If it is a single layout, *i.e.*, $|\hat{\mathcal{Y}}| = 1$ and consequently $\hat{y} = \hat{\mathcal{Y}}$ we have obtained the optimum. If it is a set of layouts, we split the set into two disjoint candidate sets $\hat{\mathcal{Y}}_1$ and $\hat{\mathcal{Y}}_2$. New bounds for those two sets are computed and denoted by $\bar{f}(\hat{\mathcal{Y}}_i)$ with $i \in \{1, 2\}$, and both candidate sets are included into the priority queue. As the bound is more tight (smaller sets), it may be that none of these candidates will be on top of the priority queue. The algorithm terminates when a single hypothesis is returned, and such hypothesis is guaranteed to be the optimum. The beauty of branch and bound is that it does not explore regions which are not promising, allowing for efficient exact inference. We refer the reader to Alg. 1 for a schematic illustration.

In order to apply branch and bound to our problem, we have to define a parameterization of the sets $\hat{\mathcal{Y}}$ as well as a bound $\bar{f}(\cdot)$ for the function of interest being $\mathbf{w}^T \phi(x, y)$. We parameterize the set of hypothesis in terms of intervals of candidate 2D ray intersections. Let $\hat{\mathcal{Y}} = \{Y_1 \cdot Y_2 \cdot Y_3 \cdot Y_4\}$ denote a set of candidate layouts defined by the product space of intervals $Y_i = [y_{i,low}, y_{i,up}]$. Note that unlike [25], these intervals are not axis aligned, but in accordance with the vanishing points. This is illustrated for an arbitrary $\hat{\mathcal{Y}}$ in Fig. 2(a) where the black rays indicate the smallest y_i in the interval, *i.e.*, $y_{i,low}$, and the red rays are drawn according to the biggest y_i in the intervals, *i.e.*, $y_{i,up}$.

We still have to derive bounds $\bar{f}$ for the original scoring function $\mathbf{w}^T \phi(x, y)$. In order for branch and bound to recover the exact solution, we need our bounds to satisfy the following two properties:



(a) Front Wall. (b) Minimal left wall. (c) Maximal left wall.

Fig. 2. (a) illustrates the max and min of the front wall in red and black respectively. (b) magenta colors the minimally possible left wall within the set of layouts bounded from below and above by the black and red rays. (c) provides the maximal left wall.

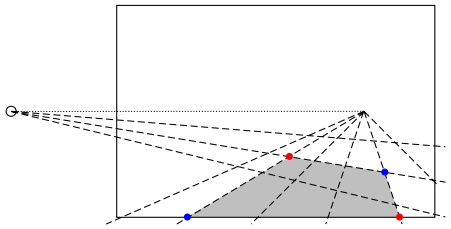


Fig. 3. Computing the content of the gray shaded area can be efficiently done in constant time using integral geometry and adding the cell contents laying to the top left of the red dotted corners while subtracting the content within the cells located to the top left of the blue dotted intersections.

1. The bound of the interval $\hat{\mathcal{Y}}$ has to upper-bound the true cost of each hypothesis $y \in \hat{\mathcal{Y}}$, *i.e.*, $\forall y \in \hat{\mathcal{Y}}$, $\bar{f}(\hat{\mathcal{Y}}) \geq \mathbf{w}^T \phi(x, y)$.
2. The bound has to be exact for every single hypothesis, *i.e.*, $\forall y \in \mathcal{Y}$, $\bar{f}(y) = \mathbf{w}^T \phi(x, y)$.

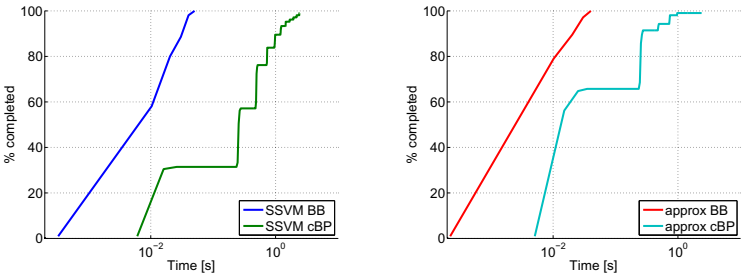
Going back to our problem, as our features $\phi_{i,\alpha,r}(x, y_\alpha)$ are always positive (*i.e.*, they represent counts), we can split the potentials in Eq. (2) into those with strictly positive weights and those with weights less or equal to zero as follows

$$\mathbf{w}^T \phi(x, y) = \sum_{\{(i,\alpha,r): i \in \{o,g\}, \alpha \in \mathcal{F}, w_{i,\alpha,r} > 0\}} w_{i,\alpha,r} \phi_{i,\alpha,r}(x, y_\alpha) + \sum_{\{(i,\alpha,r): i \in \{o,g\}, \alpha \in \mathcal{F}, w_{i,\alpha,r} \leq 0\}} w_{i,\alpha,r} \phi_{i,\alpha,r}(x, y_\alpha).$$

We can thus collapse these potentials into two functions, one that is strictly positive and one that is zero or negative by defining

$$f^+(x, y) = \sum_{\{(i,\alpha,r): i \in \{o,g\}, \alpha \in \mathcal{F}, w_{i,\alpha,r} > 0\}} w_{i,\alpha,r} \phi_{i,\alpha,r}(x, y_\alpha),$$

$$f^-(x, y) = \sum_{\{(i,\alpha,r): i \in \{o,g\}, \alpha \in \mathcal{F}, w_{i,\alpha,r} \leq 0\}} w_{i,\alpha,r} \phi_{i,\alpha,r}(x, y_\alpha).$$



(a) BB 0.01s, cBP 0.45s (b) BB 0.01s, cBP 0.15s

Fig. 4. We illustrate the run time required to obtain a fraction of completed test images on the layout data set for the proposed inference approach (BB) and a standard message passing algorithm (cBP). The results are averaged over a large set of parameters C for the two different models SSVM in (a) and approx in (b). The average times are indicated below the figures.

Inference is hence equivalently stated via the problem

$$\hat{y} = \underset{y \in \mathcal{Y}}{\operatorname{argmax}} \, f^+(x, y) + f^-(x, y), \tag{3}$$

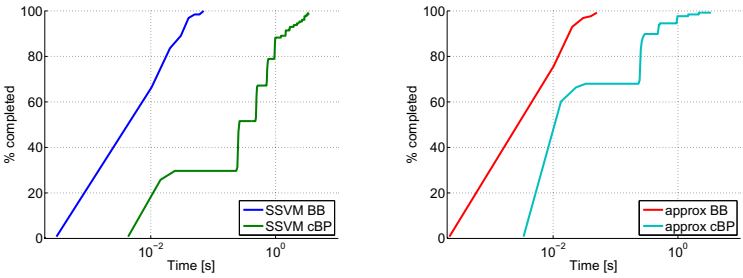
where we have used the fact that $\mathbf{w}^T \phi(x, y) = f^+(x, y) + f^-(x, y)$.

Note that in the efficient subwindow search (ESS) of Lampert *et al.* [25], the detection bounding box scoring function was also decomposed into the sum of negative and positive terms. The bounds were constructed by summing all positive terms within the outer rectangle, *i.e.*, the union of all possible members, while adding the negative terms within the interior rectangle, *i.e.*, the intersection of all possible set members. This approach is suitable if the quality function depends only on contributions from within the rectangle. With our cost function being defined on the entire image we have to find another bounding strategy. Moreover, our faces are not axis-aligned but more general convex quadrilaterals.

As our functions f^+ and f^- naturally decompose into a weighted sum over the different faces of the layout, we construct bounds by answering the question of what is the maximum positive contribution and minimum negative contribution of the score function within the set of layout candidates $\hat{\mathcal{Y}}$ for each face $\alpha \in \mathcal{F}$. The answer is simple, we need to bound each face separately by considering the minimum and maximum area that each face can take in the set. This is illustrated in Fig. 2(a) for the front wall trivially having a minimal contribution when taking $y_{i,low}$ $\forall i$ and a maximal contribution when taking $y_{i,up}$ $\forall i$. Hence we bound the contribution of the front face using

$$\bar{f}_{\textit{front-wall}}(\hat{\mathcal{Y}}) = f^+_{\textit{front-wall}}(x, y_{up}) + f^-_{\textit{front-wall}}(x, y_{low}), \tag{4}$$

where we used the subscript to restrict summation over the faces α within $f^\pm$ to the indicated set being in this case the *front-wall*. Let y_{up} and y_{low} be the 4-tuple $(y_{i,up})_{i=1}^4$ and $(y_{i,low})_{i=1}^4$. For the remaining four faces $\alpha \in \mathcal{F} \backslash \textit{front-wall}$ we need combinations of upper and lower bounding angles $y_{i,low}$ and $y_{i,up}$ to construct a



(a) BB 0.01s, cBP 0.58s (b) BB 0.01s, cBP 0.17s

Fig. 5. We illustrate the run time required to obtain a fraction of completed test images on the bedroom data set for the proposed inference approach (BB) and a standard message passing algorithm (cBP). The results are averaged over a large set of parameters C for the two different models SSVM in (a) and approx in (b). The average times are indicated below the figures.

valid function $\bar{f}_\alpha(\hat{\mathcal{Y}})$ that fulfills previously mentioned properties. We illustrate such a combination of angles for a minimal and maximal contribution of the left wall in Fig. 2(b) and Fig. 2(c). Altogether the function for scoring sets of layouts $\hat{\mathcal{Y}}$ is given by

$$\bar{f}(\hat{\mathcal{Y}}) = \sum_{\alpha \in \mathcal{F}} \bar{f}_\alpha(\hat{\mathcal{Y}}). \tag{5}$$

In order for our branch and bound inference to be practical, we need to be able to compute the bounds very efficiently. The bounds employed in [25] were efficiently computed using integral images. However, integral images are not applicable to our case, as we reason about 3D faces. Instead, we can use integral geometry [9] in order to compute these bounds in constant time.

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 [31]. In the spirit of integral images, as shown in Fig. 3, we construct 2D accumulators, each counting features (probabilities) in regions of the space defined by two rays originating from two different vanishing points. Suppose we want to compute a potential defined as counts in the gray shaded area. We can easily obtain this potential, by treating each cell as a pixel (by counting the contribution in the cell) and applying integral images to those cells. This is the concept of integral geometry introduced in [9]. Thus a potential defined in the shaded area in Fig. 3 is computed by adding the integral geometry cells to the top left of the red dotted corners while subtracting the ones to the top left of the blue dotted intersections.

4 Experimental Evaluation

We first investigate the efficiency of our method and then illustrate the state-of-the-art performance obtained by our approach via exact inference. Importantly, we are able not only to do exact inference, but also to do it faster than any other

approximate inference technique. All evaluations are carried out on two different data sets commonly used in the literature. The layout dataset [6] contains 314 images with ground truth annotation of faces, *i.e.*, left, right and front wall as well as ceiling and floor. We employed the vanishing point detector of [6], which failed in 9 training images and was successful for all test images, 105 in total. In addition, we evaluate our method on the bedroom dataset [13] which contains 309 labeled images. The data is 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. Note that the ground truth labeling is not necessarily aligned with detected vanishing points.

Efficiency: We trained our models to learn the parameters $\mathbf{w}$ using the structured support vector machine implementation of [28], denoted "SSVM," and an approximation, denoted "approx," which makes use of the same implementation but computes the cutting plane updates only approximately. Throughout this submission the stopping criteria for the message passing inference (*i.e.*, our baseline) was set to a relative duality gap of 1e-5 or a maximum of 1000 iterations. We compare the running-time of our proposed branch and bound inference technique (BB) to a message passing baseline in the form of convex belief propagation (cBP) [32]. This baseline makes use of integral geometry, and thus is equivalent to the efficient approach of [9]. Fig. 4(a) and Fig. 4(b) illustrate the percentage of completed layout test set images given a certain amount of time for models trained with "SSVM" as well as "approx" in the dataset of [6]. For the bedroom data set, we provide the results for "SSVM" training in Fig. 5(a) and for "approx" training in Fig. 5(b). To obtain meaningful statistics the given curves are averaged over a wide range of regularization tradeoff parameters $C \in \{1, 2, 5, 8, 10, 12, 15, 18, 20, 50, 100, 200, 500, 1000\}$ defined in [28]. The average run time is provided in the caption.

We note that cBP inference has a plateau-like behavior, *i.e.*, there are easy instances, but also harder ones requiring more iterations, some of them requiring the maximum number of iterations we allow. Importantly, the proposed BB approach has a roughly constant incline, and converges to the optimum faster. In contrast, the baseline is not guaranteed to get the optimum. Note that we display the results in logarithmic time scale, and thus cBP takes significantly more time in a large set of images.

To obtain quantitative results we provide the average time for inference for the layout data set and the bedroom data set in Tab. 1(a) and Tab. 1(b) respectively. The proposed BB approach outperforms the message passing technique on both data sets. Also note that the approach proposed in [24] requires about 12 minutes per image in the layout dataset.

Number of Splits: The number of iterations in branch and bound is a direct measure for the number of splits required to achieve the optimal solution. It also provides an indication for the tightness of the bounds constructed in the previous section. We emphasize that the proposed branch and bound scheme utilizes

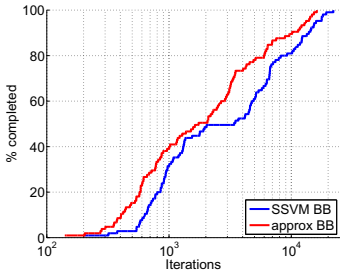
Table 1. We compare the inference run time of standard message passing (cBP) and our approach (BB) averaged over a large set of parameters C in all test instances of the (a) layout and (b) bedroom dataset for models learned with SSVM and approx.

(a) Layout data set

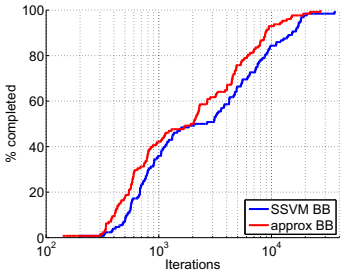
Inference \ Model	SSVM	approx
cBP	0.449 s	0.146 s
BB	**0.011 s**	**0.007 s**

(b) Bedroom data set

Inference \ Model	SSVM	approx
cBP	0.576 s	0.169 s
BB	**0.010 s**	**0.007 s**



(a) Iterations layout data set



(b) Iterations bedroom data set

Fig. 6. We illustrate the number of iterations required to obtain a fraction of completed test images on the layout data set in (a) and the bedroom data set in (b) for the proposed inference approach (BB). The results are averaged over a large set of parameters C for the two different models SSVM and approx.

integral geometry, and thus only accesses a few values on the accumulators to efficiently upper bound a set of layouts. Therefore one iteration is very fast and computable in constant time. Fig. 6(a) and Fig. 6(b) depict the cumulative distribution of the number of iterations required to obtain the optimal solution for the layout and bedroom datasets respectively. As before, we average over a large range of parameters C employed to learn our models, and compare "SSVM" and "approx." We observe that the behavior on both data sets is roughly identical, with the bedroom data set requiring on average somewhat more iterations. Independent of the way the parameters $\mathbf{w}$ are learned, we find that roughly similar number of iterations are required during inference.

Accuracy: Our branch and bound approach is faster than any previously published 3D layout estimation method. We now show its predictive performance on the layout and bedroom datasets. To this end we take the best learning settings reported in [9] and use them for our proposed BB inference. Tab. 2 and Tab. 3 show the results. Interestingly, we obtain identical results to the ones in [9]. This means that in practice cBP converges (although not guaranteed) to the optimum. Therefore we conclude that identical accuracies are obtained with our faster but most importantly, provably exact, BB inference mechanism.

Table 2. Comparison to state-of-the-art that uses the same image information on the layout data set of [6]. Pixel classification error is given in %.

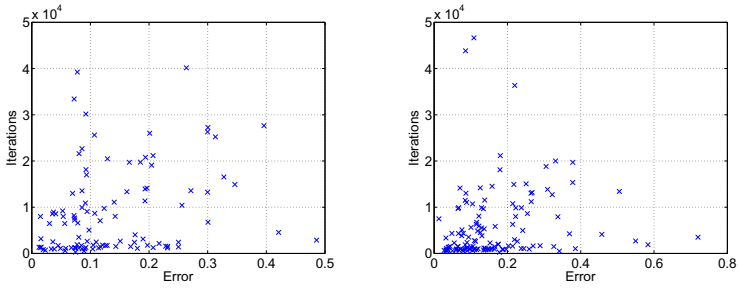
	OM	GC	OM + GC	Others	Time
[12]	-	28.9	-	-	-
[6]	-	21.2	-	-	-
[8]	22.2	-	-	-	-
[7]	24.7	22.7	18.6	-	-
[24]	-	-	-	16.3	12min
[9]	**18.64**	**15.35**	**13.59**	-	0.15s
Ours	**18.63**	**15.35**	**13.59**	-	**0.007s**

Table 3. Comparison to state-of-the-art on the bedroom data set [13]. Pixel classification error is given in %.

	[23]	[12]	[6]	[9]	Ours
Error [%]	29.59	23.04	22.94	**16.46**	**16.46**
Time [s]	-	-	-	0.17	**0.007**

Iterations vs Accuracy: Next we analyze the correlations between the test error and the number of iterations required by our BB approach. We report results for the layout data set in Fig. 7(a) and for the bedroom data set in Fig. 7(b). We observe that there is a slight but not very pronounced correlation between low errors and low number of iterations. This plot also allows reasoning about the distribution of the errors. For the layout data set we observe that the errors for a majority of the test set instances are within the interval $[0, 10\%]$. The bedroom data set seems to be more difficult with the majority of the test set instances being within the interval $[0, 20\%]$.

Visual Results: Finally, we provide visualizations of our results as well as orientation map and geometric context features for both data sets in Fig. 8 and



(a) Layout data set (b) Bedroom data set

Fig. 7. We illustrate the error for each test set instance of the layout data set in (a) and the bedroom data set in (b) with respect to the required number of iterations of the proposed BB approach

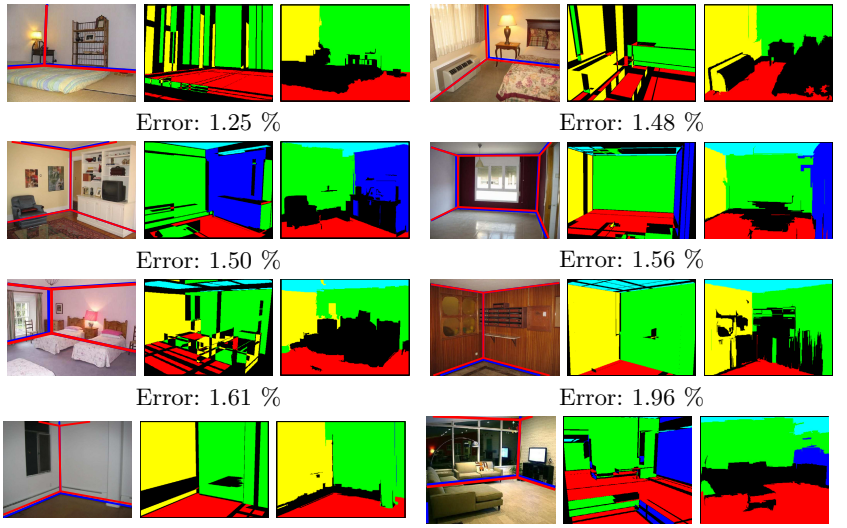


Fig. 8. The best classification results on the layout data set. The first and fourth column illustrate the image overlaid by the best possible layout obtained from ground truth labels in blue and our prediction result (red) given orientation map and geometric context features illustrated in columns 2,5 and columns 3,6 respectively.

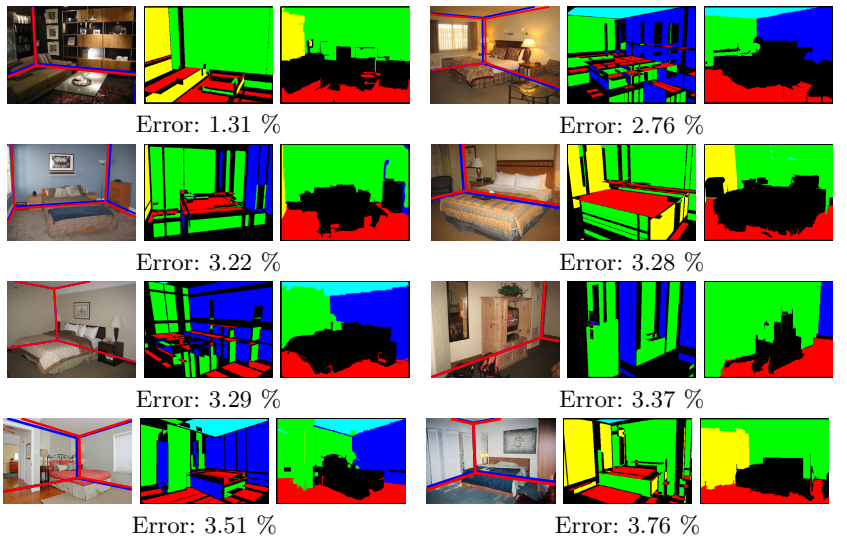


Fig. 9. The best classification results on the bedroom data set. The first and fourth column illustrate the image overlaid by the best possible layout obtained from ground truth labels in blue and our prediction result (red) given orientation map and geometric context features illustrated in columns 2,5 and columns 3,6 respectively.

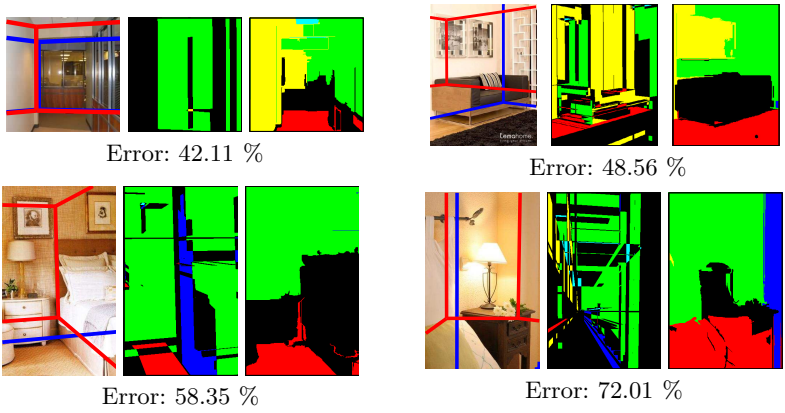


Fig. 10. The worst classification results on the layout data set (top row) and the bedroom data set (bottom row). The first and fourth column illustrate the image overlaid by the best possible layout obtained from ground truth labels in blue and our prediction result (red) given orientation map and geometric context features illustrated in columns 2,5 and columns 3,6 respectively.

Fig. 9. Here we use the model learned with OM and GC features. The blue lines overlaying the image provide the best possible discretization with 50 states while the red lines illustrate our prediction. The prediction error with respect to the pixelwise face labelings is indicated below each image. Note that the errors are due to the learned energy function as our inference is exact. Orientation maps are given in column 2 and 5 while geometric context features are depicted in column 3 and 6.

We also provide failure cases for the layout data set (top row) and the bedroom data set (bottom row) in Fig. 10. We observe two main reasons causing prediction errors. First, a failing vanishing point detection can typically not be recovered from as the rays and their configuration can be far from any true layout. The second reason for failure modes is non-informative image features due to a failing prediction in case of geometric context or wrong line detections causing misleading orientation maps.

5 Conclusion

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 presented a novel branch and bound approach which splits the label space in terms of candidate sets of 3D layouts, and bounds the energy for entire sets by constructing upper-bounding contributions of each individual face. We have employed integral geometry in order to evaluate these bounds in constant time, and show that we not only obtain the exact solution, but also in less time than approximate inference tools such as message-passing. We have demonstrated the effectiveness

of our approach in two benchmarks and show that our bounds are tight, and only a few evaluations are necessary. We plan to extend our approach to do joint inference over 3D objects as well as the layout.

References

1. Hoiem, D., Efros, A.A., Hebert, M.: Automatic Photo Pop-up. In: Siggraph (2005)
2. Saxena, A., Sun, M., Ng, A.Y.: Make3D: Learning 3D Scene Structure from a Single Still Image. PAMI (2009)
3. Gupta, A., Efros, A.A., Hebert, M.: Blocks World Revisited: Image Understanding Using Qualitative Geometry and Mechanics. In: Daniilidis, K., Maragos, P., Paragios, N. (eds.) ECCV 2010, Part IV. LNCS, vol. 6314, pp. 482–496. Springer, Heidelberg (2010)
4. Gupta, A., Satkin, S., Efros, A.A., Hebert, M.: From Scene Geometry to Human Workspace. In: Proc. CVPR (2011)
5. Geiger, A., Wojek, C., Urtasun, R.: Joint 3D Estimation of Objects and Scene Layout. In: Proc. NIPS (2011)
6. Hedau, V., Hoiem, D., Forsyth, D.: Recovering the Spatial Layout of Cluttered Rooms. In: Proc. ICCV (2009)
7. Lee, D.C., Gupta, A., Hebert, M., Kanade, T.: Estimating Spatial Layout of Rooms using Volumetric Reasoning about Objects and Surfaces. In: Proc. NIPS (2010)
8. Wang, H., Gould, S., Koller, D.: Discriminative Learning with Latent Variables for Cluttered Indoor Scene Understanding. In: Daniilidis, K., Maragos, P., Paragios, N. (eds.) ECCV 2010, Part IV. LNCS, vol. 6314, pp. 497–510. Springer, Heidelberg (2010)
9. Schwing, A.G., Hazan, T., Pollefeys, M., Urtasun, R.: Efficient Structured Prediction for 3D Indoor Scene Understanding. In: Proc. CVPR (2012)
10. Flint, A., Murray, D., Reid, I.: Manhatten Scene Understanding Using Monocular, Stereo, and 3D Features. In: Proc. ICCV (2011)
11. Lee, D.C., Hebert, M., Kanade, T.: Geometric Reasoning for Single Image Structure Recovery. In: Proc. CVPR (2009)
12. Hoiem, D., Efros, A.A., Hebert, M.: Recovering surface layout from an image. IJCV (2007)
13. Hedau, V., Hoiem, D., Forsyth, D.: Thinking Inside the Box: Using Appearance Models and Context Based on Room Geometry. In: Daniilidis, K., Maragos, P., Paragios, N. (eds.) ECCV 2010, Part VI. LNCS, vol. 6316, pp. 224–237. Springer, Heidelberg (2010)
14. Gould, S., Gao, T., Koller, D.: Region-based Segmentation and Object Detection. In: Proc. NIPS (2009)
15. Shotton, J., Winn, J., Rother, C., Criminisi, A.: TextonBoost for Image Understanding: Multi-Class Object Recognition and Segmentation by Jointly Modeling Texture, Layout, and Context. IJCV (2009)
16. Felzenszwalb, P., Girshick, R., McAllester, D., Ramanan, D.: Object Detection with Discriminatively Trained Part Based Models. PAMI (2010)
17. Hoiem, D., Efros, A.A., Hebert, M.: Putting Objects in Perspective. IJCV (2008)
18. Saxena, A., Chung, S., Ng, A.Y.: 3-D Depth Reconstruction from a Single Still Image. IJCV (2008)
19. Bao, S., Savarese, S.: Semantic Structure from Motion. In: Proc. CVPR (2011)

20. Coughlan, J., Yuille, A.: Manhattan world: Orientation and outlier detection by bayesian inference. Neural Computation (2003)
21. Coughlan, J., Yuille, A.: The manhattan world assumption: Regularities in scene statistics which enable bayesian inference. In: Proc. NIPS (2001)
22. Hedau, V., Hoiem, D., Forsyth, D.: Recovering Free Space of Indoor Scenes from a Single Image. In: Proc. CVPR (2012)
23. Pero, L., Guan, J., Brau, E., Schlecht, J., Barnard, K.: Sampling Bedrooms. In: Proc. CVPR (2011)
24. Pero, L., Bowdish, J., Fried, D., Kermgard, B., Hartley, E., Barnard, K.: Bayesian geometric modeling of indoor scenes. In: Proc. CVPR (2012)
25. Lampert, C.H., Blaschko, M.B., Hofmann, T.: Efficient Subwindow Search: A Branch and Bound Framework for Object Localization. PAMI (2009)
26. Kokkinos, I.: Rapid Deformable Object Detection using Dual-Tree Branch-and-Bound. In: Proc. NIPS (2011)
27. Taskar, B., Chatalbashev, V., Koller, D.: Learning Associative Markov Networks. In: Proc. ICML (2004)
28. Tsochantaridis, I., Joachims, T., Hofmann, T., Altun, Y.: Large margin methods for structured and interdependent output variables. JMLR (2005)
29. Lafferty, J., McCallum, A., Pereira, F.: Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In: Proc. ICML (2001)
30. Hazan, T., Urtasun, R.: A Primal-Dual Message-Passing Algorithm for Approximated Large Scale Structured Prediction. In: Proc. NIPS (2010)
31. Viola, P., Jones, M.: Rapid Object Detection using a Boosted Cascade of Simple Features. In: Proc. CVPR (2001)
32. Hazan, T., Shashua, A.: Norm-Prodcut Belief Propagation: Primal-Dual Message-Passing for LP-Relaxation and Approximate-Inference. Information Theory (2010)