Inspirited by the human vision mechanism, this paper discusses a hierarchical grammar model for 3D inference of man-made object from a single image. This model decomposes an object with two layers: (i) 3D parts (primitives) with 3D spatial relationship and (ii) 2D aspects with prediction (production) rules. Thus each object is represented by a set of co-related 3D primitives that are generated by a set of 2D aspects. The 3D relationships can be learned for each object category specifically by a discriminative boosting method, and the 2D production rules are defined according to the human visual experience. With this representation, the inference follows a data-driven Markov Chain Monte Carlo computing method in the Bayesian framework. In the experiments, we demonstrate the 3D inference results on 8 object categories and also propose a psychology analysis to evaluate our work.

Index Terms— 3D perception, man-made object, hierarchical grammar, Markov Chain Monte Carlo

1. INTRODUCTION

As reported in the psychology literature [1], the human vision system has an instinctive ability to interpret 3D information from 2D line drawing by perceiving local and global topology (i.e., object parts and spatial relations between each pair of parts). The objective of this paper is to simulate this talent of 3D perception from a single image.

3D inference and reconstruction from a single 2D intensity image is an open challenging problem in computer vision community. It has been studied from line drawing recovery with or without hidden line [2, 3, 4], by imposing strict assumptions on topology of the object. Recently, along with robust segmentation and edge detection methods arising, many researchers tried to tackle 3D object reconstruction problem from a single image [5, 6, 7, 8, 9]. Bottleneck of these approaches needs to be examined: (i) how to model 3D prior knowledge of an object, including 3D part/primitive bases (dictionary) and 3D relationship between them; (ii) how to design global inference algorithm for integrating bottom-up proposal and top-down prior. Very recently, Lin et al [8] presented a two-level iterative computation model to infer 3D structure. Han et al [9] proposed a grammar based framework and a stochastic algorithm for segmentation and reconstruction from single image. These methods show promising experimental results for tackling two mentioned points. However, to our best knowledge, there are few work dealing with complex compositional objects with quantitative analysis.

A typical compositional man-made object is shown in Fig.1(a), which consists of various parts with complex spatial relations, including supporting, occlusion, and butting, (e.g., aspect 4 in Fig.1(a)). Therefore, without hierarchical interpretation and decomposition, recovering their 3D structures by considering the input line drawings as a whole is quite difficult in such case. Furthermore, explicitly modeling the spatial constraints in hierarchical representation is also a good insight for the task.
In the object recognition research, there has been a resurgence in modeling object categories through attributed grammar and becoming popular. Representative examples are Zhu [10], Lin [11], and Dickinson [12]. These approaches show perfectly capturing both the local and global information of object category. Following this stream, this paper proposes a two-layer (2D/3D layers in Fig.1(d)) hierarchical grammar model for man-made object representation. In the 2D layer, given an input image with annotated 2D sketches, this model groups 2D geometric elements (e.g., triangles, rectangles, etc) into several 2D aspects (Fig.2(c)(d)) by using bottom-up discriminative detection and top-down generative predicting algorithms [9]. Each aspect is used as the prior to infer hidden-structures of the 3D part (Fig.2(b)), and their 2D relations are available for pursuing 3D relations through boosting learning algorithm. In the 3D layer, constrained by both the 2D spatial relations and the pursued 3D relations, several 3D parts constitute a 3D object (Fig.2(a)). These two layers are recursively computed in a Bayesian framework using data-driven Markov Chain Monte Carlo (DDMCMC) with a simulated annealing strategy.

2. PROBLEM FORMULATION

Let I be an image and S be the perfect sketch of the object. We compute the two-layer parse graph as the solution \( W = (L^{2D}, L^{3D}) \) to maximize a Bayesian posterior probability in solution space \( \Omega \). Therefore, our objective is,

\[
W^* = \arg \max_{L^{2D},L^{3D} \in \Omega} \ p(S \mid L^{3D}) p(L^{3D} \mid L^{2D}) p(L^{2D}) \quad \text{s.t.} \quad \Pi(L^{3D}) = S, \tag{1}
\]

where \( \Pi \) denotes the orthogonal projection matrix [5].

2.1. Hierarchical grammar model for 3D interpretation

The two-layer hierarchical attributed grammar model is specified by 4-tuple \( G = (V_N, V_T, \mathcal{PR}, \mathcal{R}) \). \( V_N \) and \( V_T \) are, respectively, the sets of non-leaf nodes (2D aspects/3D parts) and leaf nodes (2D elements). For deriving the source node \( O \) of the grammar, we define a set of production rules \( \mathcal{PR} \) (pyramid rule, cube rule, triangular prism rule, etc; refer to Fig.2(c)(d)), several geometric elements form an aspect using a specific production rule) according to the human visual experience. \( \mathcal{R} \) is a set of relations between each pair of non-leaf nodes,

\[
\mathcal{R} = \{ r_{ij} = < t_i, t_j, s_{r_k} > | \forall t \in V_N, s \in \mathcal{SR} \}, \tag{2}
\]

where \( \mathcal{SR} \) is a set of attributed spatial relations defined in 2D and 3D space which are illustrated at the bottom of Fig.3.

2.2. Bayesian probability models

We assume that \( A, g \) and \( P \) denote the sets of 2D aspects, 2D elements and 3D parts individually. \( K \) is the number of aspects. Consequently, there are four parts of variables in 2D layer \( L^{2D} = (K, A, g, \mathcal{R}^{2D}) \). As a non-leaf node is derived by two non-recursive production rules (e.g., scene/instance rule) in 3D layer, we use \( P \) and \( \mathcal{R}^{3D} \) to describe \( L^{3D} \). Our solution becomes

\[
W = (K, g, \{ A_i, P_{ij} \}_{i,j=0}^K, \{ r_{ij}^{2D} \}_{i=0, j=0}^K), \tag{3}
\]

where \( g(v^{2D}, E^{2D}) \) implies that each 2D element is parameterized by several vertices and edges. Attributes of \( P \) is expressed by \( (V^{3D}, E^{3D}, F^{3D}) \), where \( v^{3D}_\text{max} \) and \( v^{3D}_\text{min} \) form an axis-aligned bounding box (AABB) of the polyhedron. Furthermore, \( A \) is defined in the same way as [9].

2.2.1. Prior models — \( p(L^{2D}, L^{3D}) \)

Due to the hierarchical structure of the grammar model, we can factorize the probability \( p(L^{3D} \mid L^{2D}) \) as

\[
p(L^{3D} \mid L^{2D}) = \prod_{i=1}^{K} p(P_{i} \mid A_{i}) \prod_{j=1}^{K} p(P_{i} \mid r_{ij}^{2D}) \prod_{i,j} p(r_{ij}^{2D} \mid r_{ij}^{3D} \mid P_{i} \mid r_{ij}^{3D}). \tag{4}
\]

Firstly, topology of each 3D part is directly inferred from a specific 2D aspect and is penalized by the probability \( p(P_{i} \mid A_{i}) \) which called the local constraints. This probability takes Gibbs form and is modeled by several topological regularities such as \( SDA, WS \) and \( DP \). A modified non-symmetry measure (MNSM) is also brought in to evaluate the hidden structures of the 3D part [2]. Therefore, \( p(P_{i} \mid A_{i}) \propto \exp(-\lambda_1 SDA + \lambda_2 WS + \lambda_3 DP + \lambda_4 MNSM) \) where parameters \( \lambda_1, \ldots, \lambda_4 \) sum to 1. Secondly, 3D perception of each aspect can be guided by its corresponding 2D relations \( r_{ij}^{3D} \) which is evaluated by, a called global constraints, the probability \( p(P_{i} \mid r_{ij}^{2D}) \propto \exp(-\sum_{j=1}^{K} \lambda_j \psi_{ij}(A_i, A_j)) \). Finally, \( p(P_{i} \mid r_{ij}^{3D}) \), the 3D relative constraints, equals to

![Fig. 2. Hierarchical interpretation of the grammar model. (a) shows the compositional man-made objects. (b) indicates that several 3D parts consist an object with 3D spatial relations. (c) is a set of 2D aspects, and each aspect is generated by grouping a couple of 2D elements as illustrated in (d) by using a specific production rule.](image-url)
exp\((-\sum_{j=1}^{K} A_j \phi_{j,0}(P_i, P_j))\) and indicates that 3D relations between each pair of parts are used to refine the reconstruction results.

The energy functions \(\psi_{j,0}(\cdot)\) and \(\phi_{j,0}(\cdot)\) take quadratic form to enforce 2D and 3D relative regularities, such as ensuring that shared edges and vertices between two aspects must have similar length and shortest distance in 3D space respectively, and two parts which hold symmetric spatial relation ought to assure similar structure of their AABBS.

The prior probability for 2D layer \(p(L^{2D})\) has following form and can be modeled in the same way as \([9]\) with slightly modified \(p(L^{2D}) = p(K) \prod_{i=1}^{K} p(A_i) \prod_{ij} p(r_{ij}^{2D})\).

2.2.2. Likelihood model — \(p(S | L^{2D})\)

The likelihood probability \(p(S | L^{2D})\) is equivalent to \(p(S | g)\) and can be modeled as

\[
p(S | g) \propto \exp(-\lambda |F_S| - \sum_{g \in S} 1(g \in S)),
\]

where \(1(x) \in \{1, 0\}\) is an indicator function for Boolean variable and \(F^S\) is a set of faces of a 2D sketch.

2.3. Predicting 3D spatial relations—\(p(r_{ij}^{3D} | r_{ij}^{2D})\)

Following AdaBoost combined with a strategy of feature selection, 3D spatial relation \(r_{ij}^{3D}\) between two parts \((P_i, P_j)\) can be determined by 2D relation \(r_{ij}^{2D}\) between their corresponding aspects. We form a set of common 3D relations (e.g., adjoining, disjoining, symmetric, etc) by computing the histogram from a large data set \([14]\). All 3D relations are parametrized by the variables shown in Fig.3 bottom-right. The learning processes in quick succession are as followed

\[
H^m(x) = \sum_{t=1}^{T} a^m_t b^m_t(x), \quad G(x) = \sum_{m=1}^{M} w^m H^m(x),
\]

where \(H^m(x) : R^1 \rightarrow [-1, +1]\) and \(G(x) : R^m \rightarrow [-1, +1]\). Firstly, each feature \(f \) (Fig.3) is considered independently and modeled by its own strong classifier \(H^m(x)\). Secondly, we use all the strong feature classifiers to formulate the final ensemble \(G(x)\) within in a linear combination.

3. INFERENCER ALGORITHM

We solve 3D perception problem by maximizing the posterior probability discussed in section 2, and adopt DDMC paradigm \([15]\) within a simulated annealing strategy to search the global optima.

We design four Markov Chain dynamics. Dynamic (i) is diffusion and dynamic (ii)–(iv) are reversible jumps: (i) Randomly diffuse \(z\)-coordinates for visible vertices, and \(x\), \(y\), \(z\)-coordinates for hidden vertices of a 3D part; (ii) Add or remove spatial relation \(r_{ij}^{3D}\) between two parts; (iii) Split a visible vertex with a hidden vertex or merge two hidden vertices; (iv) Add or remove an edge between two hidden vertices.

4. EXPERIMENTS

The algorithm is implemented in C++ and 3D result is rendered by OpenGL on a PC with Core Duo 2.8GHZ CPU.

In the first experiment, we apply our method on 8 compositional object categories and the data are selected from the...
public LHI dataset [14]. 8 x 20 images (10 for each category) are tested, and consuming time for each computation is around 50 ~ 60 seconds. Fig.4 shows some representative examples that can well satisfy human visual perception.

In addition, we present a psychology experiment to further quantitatively demonstrate the 3D reconstruction accuracy. For each object category, the professional labelers first manually created 3D CAD model for each instance, and mix these 3D models with the system output results. Then a number of subjectives are asked to point out which result is system generated for each instance. Here we define a confusion rate $\delta$ to evaluate the experiment: $\delta = \text{Ave. wrong hits/total tests}$. Intuitively, we expect the $\delta$ is approaching to 50% that indicates the subjectives are confused and select randomly. The final confusion rates for all categories are reported in Tab. 1.

<table>
<thead>
<tr>
<th></th>
<th>chair</th>
<th>desk</th>
<th>monitor</th>
<th>building</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\delta(%)$</td>
<td>35.2</td>
<td>55.3</td>
<td>44.1</td>
<td>37.8</td>
</tr>
<tr>
<td></td>
<td>cupboard</td>
<td>refrigerator</td>
<td>exhauster</td>
<td>flash</td>
</tr>
<tr>
<td>$\delta(%)$</td>
<td>62.2</td>
<td>73.7</td>
<td>55.2</td>
<td>59.3</td>
</tr>
</tbody>
</table>

Table 1. Confusion rate $\delta$ for 8 object categories.

5. CONCLUSION

A two-layer grammar model for 3D interpretation of compositional man-made objects has been proposed, where the inference process follows a DDMCMC computing method in the Bayesian framework. We’ll adopt learning-based segmentation and boundary detection algorithm to compute 2D sketch and extent our work to more complex polyhedra.

6. REFERENCES