Three-dimensional Block World Reconstruction from Monocular Images by Use of an Object-based Markov Random Field Model

Hotaka Takizawa
Faculty of Engineering, Information and Systems, University of Tsukuba
1-1-1 Tennodai, Tsukuba, 305-8573, JAPAN
takizawa@cs.tsukuba.ac.jp

Daichi Tanii
Graduate School of Systems and Information Engineering,
1-1-1 Tennodai, Tsukuba, 305-8573, JAPAN

Abstract

The present report describes a novel method of reconstructing three-dimensional (3-D) scenes that include block-like objects observed in monocular images. The objects are represented by 3-D rigid models with geometrical parameters. The fidelity of these object models to the images and the consistency of relations between the object models are formulated using the Markov Random Field (MRF) model that consists of the object models. The optimal configuration of this object-based MRF model is obtained by the primal-dual interior point method. The recognition schemes are applied to actual scenes.

1. Introduction

Reconstruction of three-dimensional (3-D) block worlds from monocular images has been one of the most interesting research topics in the computer vision community. It would provide a wide range of applications such as industrial processes, video monitoring and intelligent robots. As such, extensive research has been dedicated to block world reconstruction from monocular images.

Ulrich et al. proposed a view-based approach of recognizing 3-D objects using CAD models[6]. The views of the CAD models are hierarchically arranged in a tree structure, and thus the computation time can be reduced. This method can be applied to the problems where the views can be trained beforehand. Chen et al. proposed a template-based algorithm for recognition of box-like objects in color images[1]. Box component segments are extracted from color images, and then the templates that match the segments best are sought based on the modified chamfer distances. The algorithm does not consider the relations between the objects. Gupta et al. presented a physical representation of outdoor scenes using convex block models[3]. The optimal configurations of the block models are qualitatively searched based on several physical properties, one of which is the stability of the layout of the block models in a scene. The stability is evaluated from the object densities that are estimated from an input color image by their previous method. However, the density estimation itself is not an easy problem to be solved.

The present report describes a reconstruction method of 3-D scenes that are composed of block-like objects observed in monocular images. The objects are represented by 3-D rigid models with several geometrical parameters. These object models are fitted to edges in the monocular images by evaluating (i) the fidelity of the object models to the edges and (ii) the consistency of relations between the object models. In order to evaluate them simultaneously, we introduce a Markov Random Field (MRF)[2, 7] model that is composed of the object models, which is called an *object field model*. We used the object field model for the medical image analysis[5] and the stereovision data analysis[4]. The optimal configuration of the object field model is obtained by the primal-dual interior point method based on the quasi-Newton method. We demonstrate the effectiveness of the proposed reconstruction method.

2. Object Field Model

We consider a situation where 3-D block-like objects are directly placed on a ground plane and where there are at least two observable segments crossing at the right angle on the ground plane. Based on the two segments, a virtual rectangular region is set on the

ground plane, and then it is divided into $S\ (= S_x \times S_y)$ lattice rectangles. The original and divided rectangles are called a Region Of Interest (ROI) and *cells*, respectively. The s-th cell is denoted by q_s ($s \in \mathcal{S} = \{1, 2, \cdots, S\}$). We assume that each cell includes at most one object or does not include any objects.

2.1. Object Models

In this report, boxes, cylinders and spheres are represented by rectangular solid, cylinder and sphere models, respectively. The rectangular solid model, for example, is represented by a control point $(x^{(S_R)}, y^{(S_R)})$, a rotation angle along the perpendicular axis $\theta^{(S_R)}$ and 3-D sizes $(h^{(S_R)}, w^{(S_R)}, d^{(S_R)})$. In this report, the object sizes are assumed to be known. The situation where a cell does not include any objects is represented by an *empty* model.

In q_s, the m-th object model is denoted by o_s^m ($m \in \mathcal{M}_s = \{1, 2, \cdots, M\}$). Especially, the empty model is denoted by o_s^1. The relative priori probabilities of the empty, rectangular solid, cylinder and sphere models are represented by

$$\bar{P}\left(o^{(E)}\right) = P^E, \tag{1}$$

$$\bar{P}\left(o^{(S_R)}\right) = P^{S_R}, \tag{2}$$

$$\bar{P}\left(o^{(S_C)}\right) = P^{S_C}, \tag{3}$$

$$\bar{P}\left(o^{(S_S)}\right) = P^{S_S}, \tag{4}$$

where P^E, P^{S_R}, P^{S_C} and P^{S_S} are constant values.

2.2. Belief of Object Models

We introduce another type of parameter, *belief*, which represents the degree of confidence that an object model appears in a cell. Let x_s^m denote the belief of o_s^m in q_s. x_s^m satisfies the following simplex constrains:

$$0 \le x_s^m \le 1, \tag{5}$$

$$\sum_{m \in \mathcal{M}_s} x_s^m = 1. \tag{6}$$

In Eq.(5), 1 and 0 represent the complete appearance and disappearance of o_s^m in q_s, respectively. Eq.(6) represents the exclusive appearance of object models in one cell.

3. Generation of Object Models Based on Fidelity to Edges

In each cell q_s, a list of the promising object models, $\mathcal{L}_s = \{o_s^1, o_s^2, \cdots, o_s^M\}$, is made based on the fidelity of the object models to edges in an edge image obtained by applying the Laplacian Gaussian filter to an input monocular image. Let $e \in \mathcal{E}$ denote an edge in the edge image.

The first step is to generate initial object models at several different positions in q_s, and then add them into a tentative list $\tilde{\mathcal{L}}_s$. Each object model $\tilde{o}$ in $\tilde{\mathcal{L}}_s$ is projected onto the image plane. A pixel on the line segment of the projected model is called a *projected point*, which is denoted by $g \in \mathcal{G}(\tilde{o})$. Let $e_g \in \mathcal{E}_g$ denote an edge near to g.

The posteriori probability of $\tilde{o}$ is calculated by

$$P(\tilde{o}|\mathcal{E}) = \sqrt[|\mathcal{G}(\tilde{o})|]{\prod_{g \in \mathcal{G}(\tilde{o})} \gamma(e_g^*, g)}\,, \tag{7}$$

$$e_g^* = \arg\max_{e_g \in \mathcal{E}_g} \gamma(e_g, g), \tag{8}$$

$$\gamma(e_g, g) = Sig(d(e_g, g); a_d, b_d) \cdot Sig(\theta(e_g, g); a_\theta, b_\theta), \tag{9}$$

where $Sig(x; a, b)$ is the sigmoid function with the parameters a and b $\left(\text{i.e., } Sig(x; a, b) = \frac{1}{1+\exp(a(x-b))}\right)$, $d(e, g)$ is the distance between e and g, and $\theta(e, g)$ is the difference in the edge direction between e and g. $\gamma(e, g)$ represents the fidelity of the projected point g to the edge e.

Among the object models (which are never selected) in $\tilde{\mathcal{L}}_s$, the optimal object model is selected based on the posteriori probability. By slightly changing the parameters of the selected model, new object models are generated and added to $\tilde{\mathcal{L}}_s$. The posteriori probabilities of the newly generated models are calculated.

The above selection and generation procedures are iterated until a certain number of models are generated. The $M-1$ optimal models in $\tilde{\mathcal{L}}_s$ are moved to $\mathcal{L}_s$, and, finally, the empty model is added to $\mathcal{L}_s$.

4. Formulation of State of Object Model Configuration Based on MRF

Given a set of edges $\mathcal{E}$ in the edge image, the posteriori energy function is defined by

$$U(\boldsymbol{x}|\mathcal{E}) = L(\mathcal{E}|\boldsymbol{x}) + V(\boldsymbol{x}), \tag{10}$$

where

$$\boldsymbol{x} = \left(x_1^1, x_1^2, \cdots, x_1^M, x_2^1, \cdots, x_2^M, \cdots, x_S^1, \cdots, x_S^M\right)^T \tag{11}$$

is the configuration of the object field model. $L(\mathcal{E}|\boldsymbol{x})$ is the likelihood that evaluates how the current configuration fits the edges in the edge image, and $V(\boldsymbol{x})$ is the potential energy that evaluates the consistency of cliques

of the object models. The most likely state of the object model configuration is obtained by minimizing the posteriori energy function.

The likelihood $L(\mathcal{E}|\boldsymbol{x})$ and the potential energy $V(\boldsymbol{x})$ are explained below.

4.1. Likelihood of Object Models to Edges

All the object models (except the empty models) in the lists $\mathcal{L}_1, \mathcal{L}_2, \cdots$ are projected onto the image plane as shown in Figure 1. Let $g \in \mathcal{G}$ be the projected point which comes from the m_g-th object model in the s_g-th cell (i.e., $o_{s_g}^{m_g}$). Further, let $o_{\check{s}_g}^{\check{m}_g}$ ($\check{s}_g \in \check{\mathcal{S}}_g$ and $\check{m}_g \in \check{\mathcal{M}}_{\check{s}_g}$) denote an object model that intersects with the sight segment from the camera to the origin of g.

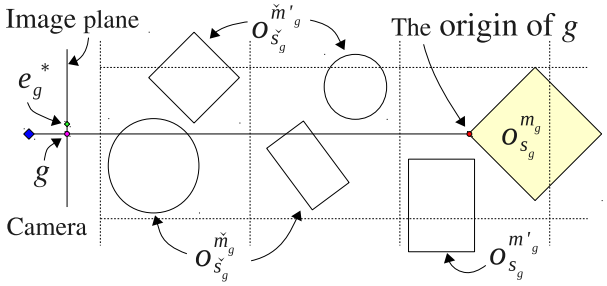


Figure 1. The relation between the object models.

The likelihood is defined by

$$L(\mathcal{E}|\boldsymbol{x}) = \frac{1}{|\mathcal{G}|} \sum_{g \in \mathcal{G}} w(g) \cdot L(\mathcal{E}|\boldsymbol{x}, g), \tag{12}$$

$$w(g) = \prod_{\check{s}_g \in \check{\mathcal{S}}_g} \left(1 - \sum_{\check{m}_g \in \check{\mathcal{M}}_{\check{s}_g}} P(o_{\check{s}_g}^{\check{m}_g}) \right), \tag{13}$$

$$\begin{aligned} L(\mathcal{E}|\boldsymbol{x}, g) &= \left(1 - \gamma(e_g^*, g)\right) x_{s_g}^{m_g} \\ &+ \sum_{m'_g \in \mathcal{M}_{s_g} \setminus m_g} \gamma(e_g^*, g)\, x_{s_g}^{m'_g} \\ &+ K_1 \sum_{\check{s}_g \in \check{\mathcal{S}}_g} \sum_{\check{m}_g \in \check{\mathcal{M}}_{\check{s}_g}} \gamma(e_g^*, g)\, x_{\check{s}_g}^{\check{m}_g} \\ &+ K_2 \sum_{\check{s}_g \in \check{\mathcal{S}}_g} \sum_{\check{m}'_g \in \mathcal{M}_{\check{s}_g} \setminus \check{\mathcal{M}}_{\check{s}_g}} \left(1 - \gamma(e_g^*, g)\right) x_{\check{s}_g}^{\check{m}'_g}, \end{aligned} \tag{14}$$

where $A \setminus B$ represents a difference set (i.e., $A \setminus B = \{x | x \in A; x \notin B\}$). K_1 and K_2 are coefficients.

The weight $w(g)$ in Eq.(13) lowers the sensitivity of the evaluation term $L(\mathcal{E}|\boldsymbol{x}, g)$ if g is occluded by the objects $\left\{ o_{\check{s}_g}^{\check{m}_g} \right\}$. $P(o_s^m)$ is the priori probability that is defined by

$$P(o_s^m) = \frac{\bar{P}(o_s^m)}{\sum_{m \in \mathcal{M}_s} \bar{P}(o_s^m)}, \tag{15}$$

where $\bar{P}(o)$ represents the relative priori probability of the object model o defined in Section 2.1.

The first term in Eq.(14) has the function of raising the belief of the object model $o_{s_g}^{m_g}$ if the fidelity $\gamma(e_g^*, g)$ is high (it implies that the object model $o_{s_g}^{m_g}$ will likely occur in the cell q_{s_g}), and vice versa. The second term, on the other hand, lowers the belief of the object models except $o_{s_g}^{m_g}$ if the fidelity is high, and vice versa.

The third term lowers the belief of the object models that intersect with the sight segment (i.e., these object models occlude $o_{s_g}^{m_g}$) if the fidelity is high, and vice versa. The fourth term raises the belief of the other models if the fidelity is high, and vice versa.

4.2. Potential Energy for Cliques of Object Models

There are two kinds of the potentials: 1-clique potential V_s^1 and 2-clique potential $V_{s,t}^2$.

The 1-clique consists of one cell q_s and its potential is defined by

$$V_s^1 = -\sum_{m \in \mathcal{M}_s} x_s^m \log P(o_s^m). \tag{16}$$

The 2-clique consists of two adjacent cells: q_s and $q_{t \in \mathcal{N}(s)}$, where $\mathcal{N}(s)$ represents a set of site indexes of the cells adjacent to q_s. The 2-cliques are classified into three categories, C_1, C_2 and C_3, based on their components as listed in Table 1.

Table 1. Three categories of the 2-cliques.

	Empty model	Substan. models
Empty model: $o^{(E)}$	C_1	C_2
Substantial models: $o^{(S_R)}$,$o^{(S_C)}$ and $o^{(S_S)}$	C_2	C_3

The potentials of these cliques are defined by the fol-

lowing non-negative function:

$$h\left(o_s^{(E)}, o_t^{(E)}\right) = h^{(E,E)}, \tag{17}$$

$$h\left(o_s^{(E)}, o_t^{(S)}\right) = h^{(E,S)}, \tag{18}$$

$$h\left(o_s^{(S)}, o_t^{(S)}\right) = -\log\left\{1 - \frac{Vol\left(o_s^{(S)} \cap o_t^{(S)}\right)}{Vol\left(o_s^{(S)} \cup o_t^{(S)}\right)}\right\}, \tag{19}$$

where $h^{(E,E)} \geq 0$ and $h^{(E,S)} \geq 0$ are constant values. The operator $Vol(\cdot)$ represents the volume of objects. The potential function $h(o_s, o_t)$ gets smaller as the object pair (o_s, o_t) is more consistent.

The 2-clique potential $V_{s,t}^2$ is defined by

$$V_{s,t}^2 = \sum_{m \in \mathcal{M}_s} \sum_{n \in \mathcal{M}_t} x_s^m \, x_t^n \, h(o_s^m, o_t^n), \tag{20}$$

and the potential $V(\boldsymbol{x})$ at the configuration $\boldsymbol{x}$ is defined by

$$V(\boldsymbol{x}) = w_1 \frac{1}{S} \sum_{s \in \mathcal{S}} V_s^1 + w_2 \frac{1}{T} \sum_{s \in \mathcal{S}} \sum_{\substack{t \in \mathcal{N}(s) \\ s<t}} V_{s,t}^2 \,, \tag{21}$$

where T is the number of the 2-cliques.

5. Minimization of Posteriori Energy

The minimization problem of the posteriori energy function of Eq.(10) under the simplex constrains of Eqs. (5) and (6) is converted into that of the following l_2 barrier penalty function:

$$\begin{aligned} F(\boldsymbol{x}; \mu) &= U(\boldsymbol{x}|\mathcal{E}) - \mu \sum_{s \in \mathcal{S}} \sum_{m \in \mathcal{M}_s} \log x_s^m \\ &+ \frac{1}{2\mu} \sum_{s \in \mathcal{S}} \left(\sum_{m \in \mathcal{M}_s} x_s^m - 1 \right)^2, \end{aligned} \tag{22}$$

from which the approximated shifted barrier KKT point is obtained by the primal-dual interior point method based on the quasi-Newton method. The obtained point represents the optimal configuration of the object field model.

6. Experimental Results

6.1. Scene 1

Figure 2(a) shows the image of an artificial block scene that includes four boxes, two cylinders and a sphere on the floor. Figure 2(b) shows the edge image with the projections of the optimal 3-D object models. Figure 2(c) and (d) show the rendering results of the object models. 5×5 cells are used which are displayed as yellow lines in the Figure 2(c) and (d).

In Figure 2(c) and (d), the positions and directions of the blocks are reconstructed almost correctly although several blocks are partially occluded by the others. Furthermore, the regions which do not include any objects in Figure 2(a) can be correctly represented by the empty models that appear as the empty cells in Figure 2(c) and (d).

It took twenty seconds to make each list of the promising object models, and twenty minutes to obtain the optimal configuration of the object field model.

6.2. Scene 2

Figure 3(a) shows the image of another scene that includes three vending machines. Figure 3(b) shows the edge image with the projections of the optimal rectangular solid models. Figure 3(c) and (d) show the rendering results of the object models. 3×3 cells are used.

The estimation of the direction of the solid model for the left-most vending machine has an error of approximately 30 degrees, which is caused by the facts that only the front surface of the vending machine can be seen from the camera and that it is partially occluded by the wall.

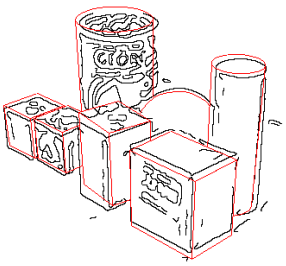
It took forty seconds to make each list of the promising object models, and sixty seconds to obtain the optimal configuration of the object field model.

7. Conclusion

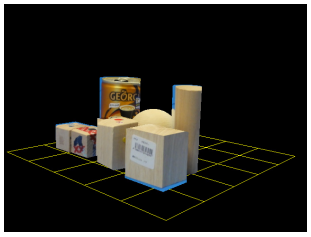
The present report describes a method of reconstructing 3-D block worlds by use of the object-based MRF model. Its optimal configuration is obtained by the primal-dual interior point method based on the quasi-Newton method. The experimental results indicate that the proposed method is promising as means of reconstructing 3-D scenes.

Our future works include the reconstruction of more complex scenes.

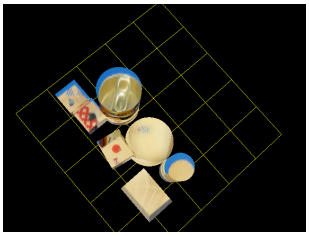
(a) Scene 1.



(b) Edges (black) and the projections of the optimal 3-D object models (red).



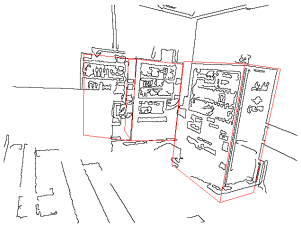
(c) The rendering result (front view).



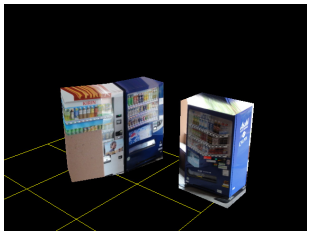
(d) The rendering result (top view).

Figure 2. Scene 1 and its reconstruction results.

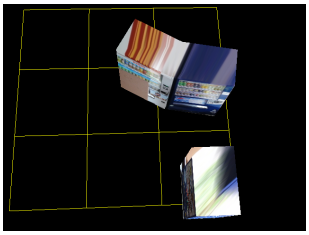
(a) Scene 2.



(b) Edges (black) and the projections of the optimal 3-D object models (red).



(c) The rendering result (front view).



(d) The rendering result (top view).

Figure 3. Scene 2 and its reconstruction results.

References

[1] C.-C. Chen and J. Aggarwal. Recognition of box-like objects by fusing cues of shape and edges. In *Proceedings of the 19 International Conference on Pattern Recognition*, pages 1–5, 2008.

[2] S. Geman and D. Geman. Stochastic relaxation, gibbs distribution, and the bayesian restoration of images. *IEEE Trans. on Pattern Analysis and Machine Intelligence(PAMI)*, PAMI-6(6):721–742, 1984.

[3] A. Gupta, A. A. Efros, and M. Hebert. Blocks world revisited: Image understanding using qualitative geometry and mechanics. In *Lecture Notes in Computer Science (European Conference on Computer Vision)*, volume 6314, pages 482–496, 2010.

[4] H. Takizawa and S. Yamamoto. Surface reconstruction from stereovision data using a 3-d mrf of discrete object models. In *18th International Conference on Pattern Recognition (ICPR2006)*, volume 1, pages 4 pages (in CD–ROM proceeding), 2006.

[5] H. Takizawa, S. Yamamoto, T. Matsumoto, Y. Tateno, T. Iinuma, and M. Matsumoto. Recognition of lung nodules from x-ray ct images using 3d markov random field models. In *Proc. of the 16th International Conference on Pattern Recognition(ICPR2002)*, volume 1, pages 10099–10102, 2002.

[6] M. Ulrich, C. Wiedemann, and C. Steger. Cad-based recognition of 3d objects in monocular images. In *Proceedings of the 2009 IEEE international conference on Robotics and Automation*, pages 1191–1198, 2009.

[7] M. D. Wheeler and K. Ikeuchi. Sensor modeling, probabilistic hypothesis generation, and robust localization for object recognition. *IEEE Trans. on Pattern Analysis and Machine Intelligence(PAMI)*, 17(3):252–265, 1995.