SCIENCE CHINA
Information Sciences

• RESEARCH PAPER •

March 2017, Vol. 60 032101:1–032101:18
doi: 10.1007/s11432-015-0710-5

Energy-based multi-view piecewise planar stereo

Wei WANG[1,2*], Lihua HU[3] & Zhanyi HU[2]

[1]*School of Network Engineering, Zhoukou Normal University, Zhoukou* 466000*, China;*
[2]*National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing* 100190*, China;*
[3]*School of Computer Science and Technology, Taiyuan University of Science and Technology, Taiyuan* 030024*, China*

Received November 30, 2015; accepted February 3, 2016; published online October 8, 2016

Abstract The piecewise planar model (PPM) is an effective means of approximating a complex scene by using planar patches to give a complete interpretation of the spatial points reconstructed from projected 2D images. The traditional piecewise planar stereo methods suffer from either a very restricted number of directions for plane detection or heavy reliance on the segmentation accuracy of superpixels. To address these issues, we propose a new multi-view piecewise planar stereo method in this paper. Our method formulates the problem of complete scene reconstruction as a multi-level energy minimization problem. To detect planes along principal directions, a novel energy formulation with pair-wise potentials is used to assign an optimal plane for each superpixel in an iterative manner, where reliable scene priors and geometric constraints are incorporated to enhance the modeling efficacy and inference efficiency. To detect non-principal-direction planes, we adopt a multi-direction plane sweeping with a restricted search space method to generate reliable candidate planes. To handle the multi-surface straddling problem of a single superpixel, a superpixel sub-segmenting scheme is proposed and a robust P^n Potts model-like higher-order potential is introduced to refine the resulting depth map. Our method is a natural integration of pixel- and superpixel-level multi-view stereos under a unified energy minimization framework. Experimental results for standard data sets and our own data sets show that our proposed method can satisfactorily handle many challenging factors (e.g., slanted surfaces and poorly textured regions) and can obtain accurate piecewise planar depth maps.

Keywords piecewise planar model, depth map, plane fitting, energy optimization, multi-view stereo

Citation Wang W, Hu L H, Hu Z Y. Energy-based multi-view piecewise planar stereo. Sci China Inf Sci, 2017, 60(3): 032101, doi: 10.1007/s11432-015-0710-5

1 Introduction

The piecewise planar model (PPM) is a widely used technique for reconstructing urban scenes, where the higher-order planarity prior significantly helps to overcome several challenging difficulties that traditional pixel-level stereos appear powerless to resolve, e.g., poorly textured regions, slanted surfaces, and inevitable occlusions. Although many excellent algorithms have been proposed in recent years, a method for automatically reconstructing a complex structured scene with completeness and accuracy remains remote.

* Corresponding author (email: wangwei@zknu.cn)

Most existing PPM methods in the literature tend to fit spatial planes along a quite restricted number of directions, or "principal directions" (e.g., the three mutually orthogonal directions in the Manhattan-world model), and hence, only geometrically simplistic models can be achieved, which are inadequate for modeling complex scene structures. Indeed, when a man-made scene is approximated by planar patches, the majority of the component planes are found in general to lie in the principal directions. However, the use of planes only along principal directions can rarely give a satisfactory approximation, and hence, in practice planes with non-principal directions must also be detected. In addition, a common practice in PPM is to assign a plane to each superpixel. Since image segmentation is a low-level process and largely based on color-homogeneity, it is inevitable that either a superpixel corresponds to several spatial planes or a spatial plane corresponds to multiple superpixels; hence, a mechanism that further sub-segments superpixel or merges them is required. To address these problems, we propose a new multi-view stereo method to reconstruct a more accurate piecewise planar depth map by incorporating scene priors and across-view photo and geometrical consistencies under a multi-level energy minimization framework, including plane-, superpixel-, and sub-superpixel-level. By virtue of such strong constraints and the power of the plane potential expression and inference, our method can efficiently detect both principal and non-principal direction planes and can also effectively handle the multi-plane straddling problem of superpixels.

The remainder of this paper is organized as follows. In Section 2, we introduce related work, followed by an overview of our method in Section 3. In Section 4, we outline two important preprocessing steps. In Section 5, we elaborate our two-stage plane inference method. In Section 6, the plane-level depth map refinement is discussed in detail. Section 7 presents our experimental results for various challenging data sets, followed by some concluding remarks in Section 8.

2 Related work

Urban scenes exhibit strong structural regularities, including poorly textured regions and multi-plane structures. The presence of such structures frequently makes it difficult for the traditional pixel-level stereo methods to obtain a complete reconstruction. In order to solve this problem, the piecewise planar assumption is usually assumed to yield a higher level reconstruction.

Most existing piecewise planar stereo methods [1–4] start from a set of candidate planes generated from reconstructed sparse or quasi-dense data (e.g., spatial points and lines), and then, use global optimization methods (e.g., Graph Cuts [5]) to infer the depth- or disparity-based planes in poorly textured regions, where the reconstructed points are either few or of very low quality. In these methods, image over-segmentation is usually adopted, because pixels of similar appearance are more likely to belong to the same plane, and then, the space patch associated with an image segment (i.e., a superpixel) is modeled as a plane.

In fact, such methods frequently heavily depend on the candidate planes used, as the initial omission of a real plane is irreversible in the subsequent optimization. To address this problem, in some studies adopted more complex segmentation schemes or similar were adopted means. For example, Bleyer et al. [6] proposed a Patch Match-based stereo method to reconstruct slanted surfaces and large untextured regions, in which various plane propagation steps are deployed to avoid the omission of real planes, yielding better results. In addition, for two-view reconstruction, some pixel- or region-matching measurements could still remain ambiguous because of fewer observations, and therefore, such methods are not effective for complex scenes.

In contrast, multi-view piecewise planar stereo could be more effective for completely reconstructing complex scenes, because of the availability of more observations and constraints. However, the traditional exhaustive plane sweeping methods [7], which directly determine the optimal planes associated with superpixels, usually lead to high computational complexity and low reliability as a result of the large size of the search space, and hence, are not practical in most cases. To solve this problem, current methods tend to align the sweeping directions to several specific scene directions. Among these methods

is Furukawa's [8], which performs the reconstruction of textureless regions of a scene by assuming the Manhattan-world model based on initial 3D-oriented points obtained from PMVS [9]. The method first obtains a set of candidate planes along three orthogonal scene directions, and then, assigns each pixel an optimal plane by pixel-wise plane labeling under the MRF framework. As a result, the method is not suitable for complex scenes with more than three scene directions. Mičušík et al. [10, 11] restricted scene directions through vanishing points and performed a superpixel-based dense reconstruction of urban scenes. However, the method can erroneously suppress a slanted plane, the normal vector of which is not consistent with the predefined main directions. Gallup et al. [12] extended the traditional plane sweeping method to account for non-fronto-parallel surfaces and performed multiple plane sweepings, where the sweeping directions were aligned to the expected surface normals of the scene. Clearly, such a method is not robust to complex scenes, because only a few sweeping directions are involved. Gallup et al. [13] also performed a piecewise planar reconstruction by incorporating high level semantic information obtained by pre-segmenting images into planar and non-planar regions through a color- and texture-based classification model. However, although such a preprocessing step does implicitly enhance the reliability of the piecewise planar assumption, it also causes the entire method to be heavily dependent on the accuracy of planar and non-planar segmentation, a difficult problem in itself.

Instead of restricting scene directions, Sinha et al. [14] first used sparse spatial points and line segments to generate candidate planes, and then, recovered a piecewise planar depth map by solving a multi-label MRF optimization problem. However, this method may incur a higher computational load in order to discover candidate planes of small size in the scene by almost exhaustive plane fitting. Even so, some real planes may still be missed because of the nature of sparse initial spatial points and line segments. Similarly, the method of Kim et al. [15] may also miss real planes, in particular for very slanted planes, since the candidate plane sampling along camera rays via only three corners of each superpixel could be unreliable because of the segmentation inaccuracy of the superpixel. In contrast to the above methods, that of Chauve et al. [16] first extracted all possible planes from unstructured spatial points using a region growing approach, and then, formulated the problem of piecewise planar reconstruction as a labeling problem of 3D space into empty or occupied regions. In fact, this method may not be robust to the noisy spatial points, as region growing can easily be entrapped in erroneous solutions. Kowdle et al. [17] proposed a piecewise planar layer-based multi-view stereo method under an energy minimization framework that combines stereo and appearance cues. This method is related to ours; however, it focuses largely on object segmentation according to the resulting piece-wise planar depth map and is not suitable for large-scale complex scenes. Recently, Bodis-Szomoru et al. [18] proposed a piecewise planar modeling method based on sparse spatial points and superpixels to generate an approximate model of the scene. In fact, although such a method is fast, it may be unreliable for plane inference, because an insufficient number of candidate planes is generated by sparser spatial points. To address this problem, this method assumes that each superpixel is sufficiently large to contain an adequate number of spatial points for plane fitting. However, such a pre-processing stage may be problematic, since a spatial patch with larger depth changes corresponding to a superpixel of large size cannot be reasonably modeled by a single plane.

3 Overview and contributions

This study focused mainly on energy-based multi-view piecewise planar stereo. Given multiple calibrated images, our goal was to recover complete piecewise planar depth maps instead of incomplete or simplistic ones reconstructed by traditional pixel-level or piecewise planar stereos. As a whole, our proposed method is a natural combination of pixel- and superpixel-level multi-view stereo and is particularly concentrated on the generation of an adequate candidate plane set and assigning an optimal plane for each space patch of the scene under the energy minimization framework. As shown in Figure 1, our method comprises three main components: preprocessing, plane inference, and depth map refinement. Each component is elaborated in the subsequent sections.

The main contributions of our work can be summarized as follows:

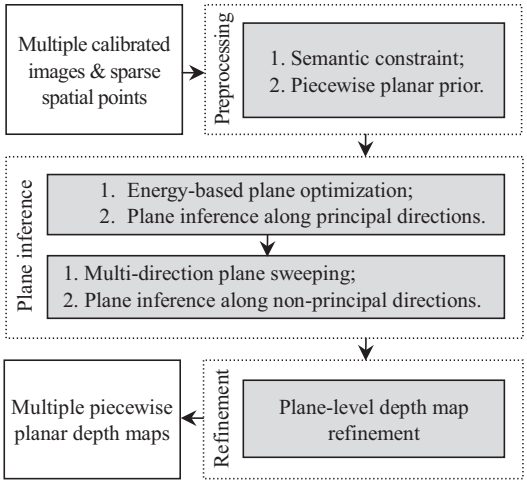


Figure 1 Flowchart of proposed method.

(1) We propose a robust candidate plane generation method for the complete reconstruction of urban scenes that includes energy-based plane optimization and plane sweeping along multiple directions, rather than merely along three orthogonal scene directions (i.e., the Manhattan-world model) as is common in the state-of-the-art algorithms [8, 11]. This allows one to recover smaller planar scene structures and obtain more accurate reconstruction results.

(2) We propose a novel and robust multi-level energy-based multi-view piecewise planar stereo method for complex scenes, which incorporates a variety of constraints and priors, such as photo-consistency, occlusion penalty, and geometric relations between spatial planes.

(3) We propose an effective plane-level depth map refinement method for which a new higher-order potential is designed to relax the hard constraint that all the pixels within a superpixel must have the same plane label, in order to generate more accurate reconstruction results.

4 Preprocessing

In this section, two important preprocessing steps are introduced: semantic constraint and piecewise planar prior.

4.1 Semantic constraint

In order to speed up the reconstruction process and reduce possible unnecessary interference by unrelated regions, such as sky and ground, we first adopt a rapid region-based semantic segmentation method [19] to extract the building regions in the current image. In fact, such a semantic segmentation also enhances the reliability of the piecewise planar assumption, because it effectively excludes those regions where the assumption is violated. Note that, for complex scenes, the semantic segmentation could also lead to inaccuracies (e.g., a small part of the sky regions is erroneously taken as a building region and vice versa) in the building region segmentation of some images. In practice, since most building regions can be correctly segmented, these inaccuracies were frequently considered negligible in this study.

4.2 Piecewise planar prior

In addition, we also use the piecewise planar priors that pixels of similar appearance are more likely to belong to the same plane and over-segment the building regions into a set of superpixels using the mean-shift image segmentation algorithm [20]. Note that a superpixel of larger size may not be modeled as a plane, since the corresponding spatial patch may straddle two or more planes. Therefore, in this study, we adopted smaller spatial and range parameters values (e.g., 3, 2) in the mean-shift algorithm to segment the current image. Clearly, considering the efficiency, these parameters values can also be

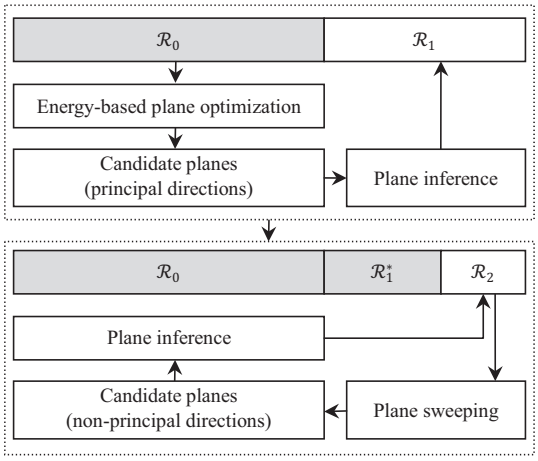


Figure 2 Robust two-stage plane inference ($\mathcal{R}_1 = \mathcal{R}_1^* \cup \mathcal{R}_2$, $\mathcal{R}_1^* \cap \mathcal{R}_2 = \emptyset$. The set of superpixels with assigned optimal planes is shaded gray).

slightly increased to reduce the number of superpixels, in particular for the segmentation of high resolution images.

5 Robust two-stage plane inference

Urban scenes contain many regularities (e.g., piecewise planarity) and usually have a restricted number of scene directions. In general, initial spatial points (or depth values) distribute mainly along several specific scene directions. Further, if a superpixel contains at least three reconstructed spatial points, the corresponding spatial plane can be fitted using RANSAC-like [21] methods and its normal is more likely to be consistent with these scene directions. For the convenience of description, here, we call these scene directions "principal directions"; our two-stage plane inference is outlined in Figure 2.

Given a reference image I_r and its k neighboring images (i.e., the images sharing a sufficient number of common visible spatial points)$\{N_i\}$ $(i = 1, \ldots, k)$, let us denote the set of all the over-segmented superpixels of building regions in I_r by $\mathcal{R}$, the set of superpixels that can be successfully modeled as planes by $\mathcal{R}_0 \subset \mathcal{R}$ and the corresponding plane set by $\mathcal{H}_0$, and the remaining superpixels by $\mathcal{R}_1 \subset \mathcal{R}$. Then, the goal of two-stage plane inference is to assign an optimal plane to each superpixel $s \in \mathcal{R}$.

Note that a plane in $\mathcal{H}_0$ obtained by RANSAC-style methods could be false because of the fitting of some unreliable spatial points and needs to be eliminated to enhance the reliability of the plane inference (see Subsection 5.3). Moreover, in order to obtain a complete depth map, each superpixel in $\mathcal{R}_1$ also needs to be assigned an optimal plane along either the principal or non-principal directions. As shown in Figure 2, in this study, the problem was solved by performing robust two-stage plane inference along the principal and non-principal directions, respectively.

In the next sections, we first introduce the robust superpixel-based energy formulation and two candidate plane generation methods for the two-stage plane inference, and then, describe the overall process of the plane inference.

5.1 Superpixel-based energy formulation

For the problem of plane inference, the energy function in our energy formulation is defined on the set of superpixels $\mathcal{R}$; the associated label set $\mathcal{L}$ is the indices of the candidate planes. Let $f_s \in \mathcal{L}$ denote the current label assigned to superpixel $s \in \mathcal{R}$. Now, our goal is to find the optimal solution f such that the energy function $E(f)$ defined in (1) is minimized:

$$E(f) = \sum_{s \in \mathcal{R}} E_{\text{data}}(s, f_s) + \lambda_{\text{smo}} \cdot \sum_{t \in N(s)} E_{\text{smooth}}(f_s, f_t), \tag{1}$$

where λ_{smo} is the weight of the smoothness term and $N(s)$ denotes all the neighboring superpixels of the superpixel s.

Next, we elaborate the data term $E_{\text{data}}(s, f_s)$ and pair-wise smoothness term $E_{\text{smooth}}(f_s, f_t)$.

5.1.1 *Data term*

The data term is comprised of three components: the photo-consistency measure $E_{\text{pho}}(s, f_s)$, geometric constraint $E_{\text{geo}}(s, f_s)$ and plane prior. First, we define the basic data term consisting of $E_{\text{pho}}(s, f_s)$ and $E_{\text{geo}}(s, f_s)$ as

$$E'_{\text{data}}(s, f_s) = \kappa \cdot E_{\text{pho}}(s, f_s) + (1 - \kappa) \cdot E_{\text{geo}}(s, f_s), \tag{2}$$

where the constant κ is used to adjust the weight of the geometric constraints against the photo-consistency measure.

(1) Photo-consistency measure

Let D_r and $\{D_i\}\,(i = 1, \ldots, k)$ respectively denote the initial depth maps corresponding to image I_r and $\{N_i\}\,(i = 1, \ldots, k)$. For a superpixel $s \in \mathcal{R}$, if the cost of a single pixel $p \in s$ with respect to a neighboring image N_i and a plane H is $C_s(p, H, N_i)$, then $E_{\text{pho}}(s, f_s)$ is simply the sum of all such costs over all the neighboring images and over all the pixels belonging to s. Therefore, we first give a definition of the cost, $C_s(p, H, N_i)$.

Given $p \in s$, let P denote the intersection point of its back-projection ray with the plane H_s corresponding to the label f_s, M the set of the pixels with the reconstructed depth value in depth map D_i, and $\overline{M}$ the set of pixels without depth value. In addition, let $H(p)$ denote the projection point of P in neighboring image N_i induced by the plane H_s and $d(H(p))$ the depth value of P with associated image N_i.

Then the cost $C_s(p, H, N_i)$ is defined as

$$C_s(p, H_s, N_i) = \begin{cases} \min\left(\left\|I_r(p) - N_i\left(H_s(p)\right)\right\|, \delta\right), & H_s(p) \in \overline{M}, \\ \lambda_{\text{occ}}, & D_i\left(H_s(p)\right) \leqslant d\left(H_s(p)\right), \\ \lambda_{\text{err}}, & D_i\left(H_s(p)\right) > d\left(H_s(p)\right), \end{cases} \tag{3}$$

where $\left\|I_r(p) - N_i\left(H_s(p)\right)\right\|$ denotes the absolute difference of the normalized color (i.e., the value is between 0 and 1) at $p \in I_r$ and $H_s(p) \in N_i$, and the parameter δ (set to 0.5 in this paper) is a truncation threshold to address the robustness concern related to occlusion regions. The constants λ_{occ} and λ_{err} are respectively the occlusion penalty and the free-space violation penalty.

In (3), the first case is meant that if $H_s(p)$ is in region $\overline{M}$, plane H_s is more likely to be a real plane, and the photo-consistency cost is measured by the dissimilarity of color distribution. The second case means that if P is occluded by a reliable spatial point (i.e., the spatial point has been verified as a reliable one in the reconstruction process) with depth value $D_i(H_s(p))$, a penalty constant λ_{occ} is assigned. The third case indicates that if P occludes a reliable spatial point with depth value $D_i(H_s(p))$, a larger penalty is needed, because a reliable spatial point is unlikely to be occluded. In this case, λ_{err} should be larger than λ_{occ}.

Thus, the photo-consistency measure $E_{\text{pho}}(s, f_s)$ is simply defined as

$$E_{\text{pho}}(s, f_s) = \frac{1}{k \cdot |s|} \sum_{i=1}^{k} \sum_{p \in s} C_s(p, H_s, N_i). \tag{4}$$

Here, the label f_s is the index of the plane H_s and $|s|$ is the number of pixels in superpixel s.

(2) Geometric constraint

In practice, if we have obtained some spatial points, the projections of which are within superpixel s, the optimal plane associated with superpixel s can be reliably determined from the known candidate planes. Clearly, the candidate plane that is closer to these spatial points is more likely to be the real

plane associated with superpixel s; then, we formulate the geometric constraints as

$$E_{\text{geo}}(s, f_s) = \begin{cases} \lambda_{\text{dis}}, & |P_s| = 0, \\ \frac{1}{1+\mathrm{e}^{-d(P_s, H_s)}}, & |P_s| > 0, \end{cases} \tag{5}$$

where P_s is the set of known spatial points, $d(P_s, H_s)$ the average distance of points in P_s to the plane H_s, and λ_{dis} the distance penalty.

In fact, for superpixel s, the optimal plane is usually more likely to be the plane with the highest prior, unless the corresponding local evidence (e.g., the photo-consistency cost) strongly disagrees. Here, the plane prior $p(H_s)$ is defined as the ratio of the number of the spatial points belonging to the plane H_s to the total number of the spatial points belonging to all the planes along this scene direction. Note that, at the initial stage, each superpixel defines its own plane; the corresponding plane prior is simply set to 1.

Then, as in the method presented in [12], the data term with the plane prior is

$$E_{\text{data}}(s, f_s) \propto E'_{\text{data}}(s, f_s) - 2\sigma^2 \log\left(p(H_s)\right), \tag{6}$$

where the scalar σ is set empirically.

Note that, in the process of the two-stage plane inference, $p(H_s)$ in (6) is automatically updated within each iteration, since more planes are reliably determined along either the principal or non-principal directions. That is, $p(H_s)$ becomes increasingly important to the geometric constraints and the photo-consistency measure. In practice, this usually allows the plane inference process to converge faster.

5.1.2 *Smoothness term*

The pair-wise smoothness term we used is defined as

$$E_{\text{smooth}}(f_s, f_t) = \omega_{st} \cdot \delta(f_s \neq f_t), \tag{7}$$

where $\delta(x)$ is the indicator function that equals 1 if x is true, and 0 otherwise, and ω_{st} is a discrepancy penalty for neighboring superpixels s and t.

According to the piecewise planar assumption, the smoothness term should encourage neighboring superpixels to take the same plane label or the parameters of the planes should vary smoothly between similar colored regions. Therefore, ω_{st} should be adjusted adaptively according to the color dissimilarity and the length of the boundary between neighboring superpixels; it is defined as

$$\omega_{st} = b_{st} \cdot (1 - \|c(s) - c(t)\|) \cdot \|n(s) - n(t)\|, \tag{8}$$

where $c(s)$ is the mean color normalized to a range of 0–1 of superpixel s, $n(s)$ is the unit normal vector of the plane H_s corresponding to the label f_s, $\|\cdot\|$ could be any error measure, and $b_s t$ is the length of the shared boundary between superpixels s and t, divided by the shorter superpixel circumference.

Clearly, if the color difference between two superpixels is smaller and the boundary between them is longer, the corresponding smoothness constraint should be stronger.

5.2 Candidate plane generation

In this section, we focus on the generation of suitable candidate planes for the two-stage plane inference first by energy-based plane optimization and then by multi-direction plane sweeping. Note that our plane sweeping method is based on the resultant depth map of the first-stage plane inference, and hence, its search range is much reduced, despite its inherently exhaustive search nature.

5.2.1 *Energy-based plane optimization for principal-direction planes*

In fact, not only does $\mathcal{H}_0$ usually contain some outliers, but also a large proportion of these are nearly co-planar because of image over-segmentation. Hence, it is desirable to fuse slightly different planes into a globally optimal one under a certain criterion. In addition, as shown in Figure 2, the principal directions also need first to be determined from $\mathcal{H}_0$ to facilitate the subsequent plane inference. In other words, a small number of optimal planes that best explain the initial structures of the scene needs to be selected from $\mathcal{H}_0$ to act as the label set to perform the first-stage plane inference for superpixels in $\mathcal{R}_1$.

Conceptually speaking, this problem can also be solved by applying existing multi-model fitting methods, such as PEaRL [22, 23] or RCMSA [24] directly over initial spatial points or depth maps. However, such methods usually tend to miss many real planes, since they are primarily concerned with fitting sparse spatial points.

Inspired by the PEaRL algorithm, here, we formulate the problem of extracting principal-direction planes as a plane labeling problem under an energy minimization framework. Since our method is an improvement on PEaRL, we first summarize the main steps of PEaRL as follows.

(1) Propose: randomly sample data points (e.g., spatial points) to generate initial candidate models (e.g., planes).

(2) Expand: perform model optimization based on initial or pruned candidate models via α-expansion [5] for a predefined energy function. If the energy does not decrease, then stop.

(3) Re-estimate: prune candidate models and obtain a new set of models; then go to Step (2).

In practice, the standard PEaRL algorithm may be problematic for the following reasons.

(1) The spatial coherence assumption, i.e., neighboring data points correspond approximately to the same model, is not always valid, because the neighbors computed by Delaunay triangulations are not always true spatial neighbors, which usually leads to erroneous results caused by the smoothness constraints in the energy function.

(2) In PEaRL, a larger set of model proposals usually leads to better solutions, since it is more likely to contain good model proposals. However, randomly sampling data points to generate model proposals frequently incur a higher computational load and could also miss some likely models if the number of sampling is small.

(3) Outliers in the model proposals at each iteration may reduce the reliability of the next "Expand" and "Re-estimate" step.

In this study, the above three points were addressed by formulating the plane labeling problem directly over the set of superpixels $\mathcal{R}_0$. Specifically, let $\mathcal{L}$ denote the label set that is originally constructed by the indices of $\mathcal{H}_0$ and pruned in the "Re-estimate" step. Currently, our goal is to assign each superpixel $s \in \mathcal{R}_0$ a label $f_s \in \mathcal{L}$ such that the energy function $E(f)$ defined in (9) is minimized:

$$E(f) = \sum_{s \in \mathcal{R}_0} E_{\text{data}}(s, f_s) + \lambda_{\text{smo}} \cdot \sum_{t \in N(s)} E_{\text{smooth}}(f_s, f_t) + \lambda_{\text{lab}} \cdot \sum_{L \in \mathcal{L}} E_{\text{label}}(f). \tag{9}$$

Here, λ_{smo} and λ_{lab} are respectively the weight of the smoothness term and label term.

In (9), the data term $E_{\text{data}}(s, f_s)$ encodes the geometric errors caused by assigning label f_s to a superpixel s and is defined as

$$E_{\text{data}}(s, f_s) = d(P_s, H_s), \tag{10}$$

where $d(P_s, H_s)$ is defined in (5).

The smoothness term $E_{\text{smooth}}(f_s, f_t)$ is the same as that in (7).

The label term $E_{\text{label}}(f)$ is used to penalize the number of labels in the solution f in order to obtain a small number of planes that can best explain the scene structure. It is defined as

$$E_{\text{label}}(f) = \mathrm{e}^{-\lambda_0 \cdot |L|} \cdot \delta_L(f), \tag{11}$$

where $|L|$ is the number of superpixels assigned to label L after the "Expand" step and the model complexity penalty λ_0 is a constant and selected empirically. The indicator function $\delta_L(\cdot)$ is defined on

label subset $L \in \mathcal{L}$ as

$$\delta_L(f) = \begin{cases} 1, \exists s : f_s = L, \\ 0, \text{ otherwise.} \end{cases} \tag{12}$$

As compared to the standard PEaRL algorithm, our improved energy formulation has the following three advantages:

(1) A reliable spatial neighborhood system and adaptive smoothness constraints can be constructed effectively, which helps to enhance the reliability of the plane optimization;

(2) The planes in $\mathcal{H}_0$ are more reliable, since a superpixel with homogenous color distribution is more likely to be associated with a real plane;

(3) The number of plane proposals is not excessively large, which helps to enhance the computational efficiency of the "Expand" and "Re-estimate" steps.

In general, after the current "Expand" step, correct labels (i.e., planes) can frequently find more spatially coherent inliers (i.e., superpixels [1)]), while erroneous labels obtain fewer or no inliers. In addition, the erroneous labels can be effectively detected and further assigned to the optimal inliers in the process of the two-stage plane inference (see Subsection 5.3).

5.2.2 *Multi-direction plane sweeping for non-principal-direction planes*

After the plane inference along the principal directions, many superpixels in $\mathcal{R}_1$ can be assigned a reliable plane, as shown in Figure 2. These superpixels are denoted by $\mathcal{R}_1^* \subseteq \mathcal{R}_1$ and the remaining superpixels by $\mathcal{R}_2 \subseteq \mathcal{R}_1$. Evidently, the plane corresponding to superpixel $s \in \mathcal{R}_2$ needs to be further inferred. However, the planes along non-principal directions cannot easily be determined because of various uncertainties, and therefore, we resort to a plane sweeping approach to generate candidate planes for the second-stage plane inference. Conventional multi-view plane-sweep-based stereo methods sweep planes either exhaustively over all the possible scene directions or only along several specific directions. Clearly, exhaustive sweeping would lead to a larger number of candidate planes, and hence higher computational complexity. However, very constrained sweeping is not suitable for complex scenes. In this study, the exhaustive plane sweeping could be effectively simplified, because the solution space of the plane inference problem for superpixels in $\mathcal{R}_2$ has been dramatically reduced by the known planes corresponding to superpixels in $\mathcal{R}_0 \cup \mathcal{R}_1^*$.

In practice, for a candidate plane H_s of superpixel $s \in \mathcal{R}_2$ in the reference image, the plane H_s could be considered unreliable if the photo-consistency measure $E_{\text{pho}}(s, f_s)$ is too large. Therefore, we sample the hemisphere of all visible surface normals and compute the $E_{\text{pho}}(s, f_s)$ value of each plane along each scene direction. Finally, the most probable K (set to 10 in this paper) planes with smaller $E_{\text{pho}}(s, f_s)$ values are selected from those planes for each superpixel s, and all such selected planes for all the superpixels in $\mathcal{R}_2$ are used to constitute the label set for the second-stage plane inference.

5.3 The plane inference process

Our two-stage plane inference is performed first along the principal directions and then along the non-principal directions. In each stage, in order to enhance the robustness of the plane inference, those superpixels that are assigned to unreliable planes are detected in time and further inferred. Therefore, an iterative manner is adopted.

In each iteration, for a superpixel s, the percentage ρ of the number of pixels with reliable depth values to the total number of pixels is first computed after eliminating the outliers of the current depth map based on the depth consistency measure [25]. Then, if the percentage ρ is smaller than a predefined threshold ϑ, the corresponding plane assigned to s is considered unreliable and is updated in the next iteration. Similarly to that for the reference image, the plane inference is also performed over its multiple neighboring images (e.g., left and right neighboring images) in order to detect unreliable planes.

Moreover, at each stage, the iterative process should stop when no superpixel is updated. To achieve this, we store those superpixels that are not assigned reliable planes at stage 1 to r_1, and those at

1) A spatial plane can be associated to a number of superpixels in the reference image. If this plane is the real plane corresponding to a superpixel, the superpixel is called an inlier of this plane, and otherwise an outlier.

stage 2 to r_2. Accordingly, the number of such superpixels during the ith iteration is $\left|r_1^i\right|$ or $\left|r_2^i\right|$, and then, the corresponding plane inference stops if $\left|r_1^i\right|$ or $\left|r_2^i\right|$ no longer change at two successive iterations. In addition, we also store the set of the optimal planes produced in the two-stage plane inference to $\mathcal{H}_{\text{dmo}}$, which is used in the subsequent plane-level depth map refinement.

The process of the plane inference is summarized in Algorithm 1.

Algorithm 1 Robust two-stage plane inference

Input: initial depth map $\mathcal{D}_0$;
Output: updated depth map $\mathcal{D}$ and reliable plane set $\mathcal{H}_{\text{dmo}}$;
Initialization: $r_1 = \phi$, $r_2 = \phi$, and $\mathcal{D} = \mathcal{D}_0$;
% of the plane inference along principal directions
1: Generate candidate plane set along principal scene directions (see Subsubsection 5.2.1);
2: **While** $r_1 = \phi$ or $\left|r_1^{i+1}\right| - \left|r_1^i\right| > 0$
3: Solve (1) using Graph Cuts;
4: Update $\mathcal{D}$, detect unreliable planes, and add the corresponding superpixels to r_1;
5: Update the priors of all candidate planes;
6: **Endwhile**
7: Save all the reliable planes to $\mathcal{H}_{\text{dmo}}$.
% of the plane inference along non-principal directions
8: Generate candidate plane set along non-principal scene directions (see Subsubsection 5.2.2);
9: **While** $r_2 = \phi$ or $\left|r_2^{i+1}\right| - \left|r_2^i\right| > 0$
10: Solve (1) using Graph Cuts;
11: Update $\mathcal{D}$, detect unreliable planes, and add the corresponding superpixels to r_1;
12: Update the priors of all candidate planes;
13: **Endwhile**
14: Save all the reliable planes to $\mathcal{H}_{\text{dmo}}$.

In fact, at the first stage, since most of the initial planes are distributed along principal directions and the plane priors are incorporated into the energy function, more planes associated with superpixels in $\mathcal{R}_1$ can be reliably determined. Then, the corresponding constraints produced by the planes along the principal directions become tighter, and thus, the performance of the exhaustive plane sweeping and the plane inference along non-principal directions at the second stage is fast. Moreover, all the superpixels with $\rho > \vartheta$ are considered as constants in iterations. Hence, Algorithm 1 can always converge rapidly.

6 Plane-level depth map refinement

After the two-stage plane inference, each superpixel of $\mathcal{R}$ is usually assigned a relatively reliable plane. Unfortunately, in complex scenes, some over-segmented superpixels could straddle two or more planes with larger depth changes and cannot be reasonably modeled by single planes. Therefore, the planes associated with such superpixels could be unreliable, and usually have a smaller ρ value (see Subsection 5.3). To address this issue, an effective plane-level depth map refinement process is performed via incorporation of the robust P^n Potts potentials.

More specifically, for superpixel s, if the ρ value is small, the associated plane is regarded as unreliable, and then, the pixels in s are classified into two sets, one with reliable depth values denoted as s_0 and the second without depth values denoted as $\overline{s_0}$. Then, we consider s_0 belongs to a plane, but $\overline{s_0}$ is further segmented according to their color similarity by an affinity propagation (AP) algorithm [26]. Further, we eliminate small sub-superpixels by merging them with other sub-superpixels according to the color similarity, since a larger sub-superpixel usually has a better distinctiveness than a smaller one. In fact, such a merging process also reduces the number of variables in the higher-order potential, which helps to speed up the subsequent energy minimization. Here, let $\{s_i\}\,(i = 0, 1, \ldots, m)$ denote the set of partitioned sub-superpixels. Clearly, the more reliable plane for sub-superpixel s_i should be further inferred such that the reconstructed depth values of pixels belonging to s_i also satisfy the depth consistency measurement [25]. In this work, we adopt the robust P^n Potts model [27] to solve this

problem. The robust higher-order potential is defined as

$$E_{\text{higher}}(s) = \begin{cases} \frac{\rho}{Q} \cdot G(s), & \text{if } \rho \leqslant Q, \\ G(s), & \text{otherwise,} \end{cases} \tag{13}$$

where Q is the truncation parameter, which controls the rigidity of the higher order potential, and $G(s)$ is a quality measurement of superpixel s, as

$$G(s) = \alpha \cdot \exp\left(-\frac{\sum_{s_i \in s} \left(c(s_i) - \mu \right)^2}{\beta \cdot m} \right), \tag{14}$$

where $\mu = \sum_{s_i \in s} c(s_i)/m$, $c(s_i)$ is the average color of sub-superpixel s_i, and the scalars α, β adjust the magnitude and sharpness of $G(s)$.

Obviously, the smaller the color difference between $s_i \in s$ and $s_j \in s$, the greater is the chance of s being a true superpixel, and the planes H_{s_i} and H_{s_j} are more likely to be the same one, which should impose a larger penalty on superpixel s. Otherwise, a smaller penalty should be imposed to encourage the difference of H_{s_i} from H_{s_j} or H_s. On the other hand, the smaller ρ is, the lower is the reliability of the superpixel, and the robust P^n Potts model imposes a smaller penalty on it such that each sub-superpixel s_i is encouraged to take a plane other than H_s. Given ρ, if the truncation parameter Q is larger, the penalty is relatively small, and then, sub-superpixel s_i is more likely to be assigned to a plane other than H_s.

Finally, let $\mathcal{R}_{\text{dmo}}$ denote the union set of all the good quality superpixels with $\rho > \vartheta$ and all the sub-superpixels. Then, our energy function is formulated over $\mathcal{R}_{\text{dmo}}$ by incorporating the higher-order potentials $E_{\text{higher}}(s)$ into the pair-wise energy function defined in (1) as

$$E(f) = \sum_{s \in \mathcal{R}_{\text{dmo}}} E_{\text{data}}(s, f_s) + \lambda_{\text{smo}} \cdot \sum_{t \in N(s)} E_{\text{smooth}}(f_s, f_t) + \lambda_{\text{hig}} \cdot \sum_{s \in \mathcal{R}_{\text{dmo}}} E_{\text{higher}}(s), \tag{15}$$

where λ_{hig} is the weight of the higher-order term.

According to the method in [27], the robust P^n Potts model can be transformed to a pair-wise energy by incorporating a small number of additional auxiliary nodes, and minimized by α-expansion. Note that all the superpixels with $\rho > \vartheta$ and sub-superpixels with reliable depth values are considered as constants in this step, and the current label set is the indices of planes associated with $\mathcal{H}_{\text{dmo}}$ obtained by the previous two-stage plane inference. Moreover, we found empirically that a single minimization for (15) is sufficient, and no iterations as in the two-stage plane inference are needed.

7 Experimental results

To evaluate the performance of our method, we conducted experiments on several data sets of scenes in which planar structures dominate: (1) the Oxford VGG data set [2)]: the Valbonne images (768×512), the Wadham images (1024×768), and the Merton images (1024×768); (2) our own data sets: the Tsinghua physics building (TPB) images (819×546), the life science building (LSB) images (1092×728), and the CITY images (1224×1848).

In addition, we note that, for all our experiments, only the images to the left and right of the reference image were taken as its neighboring images for its depth map reconstruction, i.e., k=2 in (4). Clearly, more neighboring images could enhance the reliability of the whole algorithm, but would also incur a higher computational load. Figure 3 shows two sample images of each data set.

In order to quantitatively evaluate our method, we assessed the depth consistency, as in [25]. Specifically, let D_r and $\{D_i\}\,(i = 1, \ldots, k)$ denote the depth map corresponding respectively to the reference image I_r and its neighboring images $\{N_i\}\,(i = 1, \ldots, k)$. Given $m \in D_r$ and its corresponding spatial point X_m, let $d(X_m, D_i)$ denote the depth value of X_m in D_i, and $\lambda(X_m, D_i)$ the depth value computed

2) http://www.robots.ox.ac.uk/ vgg/data/data-mview.html.

Figure 3 Two sample images of the data sets (the images in the first row are the reference images). (a) Valbonne; (b) Wadham; (c) Merton; (d) TPB; (e) LSB; (f) CITY.

by X_m under the camera parameters of D_i. Then, X_m is considered reliable with respect to D_i if the difference between $d(X_m, D_i)$ and $\lambda(X_m, D_i)$ is smaller than a predefined threshold ε. This reliability should be reinforced by considering N (set to 2 in this study) neighboring depth maps. Therefore, we defined the reconstruction accuracy T as

$$T = \frac{1}{|A|} \sum_{m \in A} \delta \left(\sum_{i=1}^{k} \delta \left(\frac{\|\lambda(X_m, D_i) - d(X_m, D_i)\|}{d(X_m, D_i)} < \varepsilon \right) > N \right), \tag{16}$$

where A denotes the set of pixels in building regions.

Moreover, we also used the number of main scene planes to evaluate the corresponding reconstruction accuracy.

All the experiments were conducted on a desktop PC with Intel Core2Duo 3.2 GHz CPU and 16 GB RAM, and each algorithm in all experiments was implemented in parallel C++.

7.1 Parameter settings

In all the experiments, we adopted the same parameter setting. The proposed method seemed not sensitive to parameters settings and most of the parameters were fixed. Specifically, our results showed that our method is effective with the following settings: κ=0.5, λ_{occ}=2, λ_{err}=5, λ_{dis}=2, σ=5, λ_0=50, ϑ=0.6, λ_{smo}=0.6, λ_{lab}=0.2, λ_{hig}=0.2, α=5, and β=15. Next, we give some explanations for the parameter setting.

In practice, κ can be set according to the number of initial spatial points to adjust the weight of the geometric constraints against the photo-consistency measure. κ=0.5 means the two parts are equally important for computing the data term. In addition, the free-space violation λ_{err} should be larger than the occlusion penalty λ_{occ} as discussed in Subsection 5.1, and thus, they were set to 5 and 2, respectively, with respect to the color dissimilarity measure. Similarly, in (5), the distance penalty should be larger than the maximum of the average distance (i.e., 1), and thus, was set to 2. Moreover, σ and λ_0 were respectively set to 5 and 50 empirically according to the characteristics and complexity of the scene structure. The threshold ϑ was used to detect outliers (e.g., unreliable planes) in the process of the two-stage plane inference. A larger ϑ helps to filter out more outliers to achieve better results in the next iterations, and also leads to a higher computational load. Here, we set it to 0.6.

In addition, a larger value of λ_{smo} could enhance the smoothness of neighboring planes, but give rise to the risk of some similar neighboring planes being combined into a single one. Likewise, a larger value of λ_{lab} also helps to reduce the plane number in the process of extracting principal-direction planes. However, too large a value of λ_{smo} and λ_{lab} could lead to the loss of scene details and an insufficient number of candidate planes. Our results show that the proposed method performed well for all the datasets in our experiments when λ_{smo}=0.6 and λ_{lab}=0.2. As for the remaining parameters in the high-order energy term, although they could be estimated through learning from training data, here we simply determined them in an enumerating manner that traverses all possible settings in the predefined range, i.e., $\lambda_{\text{hig}} \in [0, 1]$, $\alpha \in [1, 20]$, $\beta \in [1, 20]$ and the ratio $Q/\rho \in [1, 10]$. Finally, we took the parameter settings with the best reconstruction accuracy T as the optimal ones. Further, our results showed that

Figure 4 Reconstruction process of our method for Valbonne data set. (a) Initial sparse spatial points; (b) superpixels with initial noisy planes and textured noisy planes; (c) candidate planes along principal directions (different gray denotes different plane); (d) final plane reconstruction; (e) textured reconstruction; (f) top view of the textured reconstruction.

these parameter settings are almost similar for all the datasets, and therefore, they were fixed for all the datasets. As an example, we report some experiments on the setting of Q which appears relatively more important for the improvement of the final reconstruction accuracy. Note that the enumerating process is faster when performed only in the stage of the depth map refinement.

7.2 Results and analysis

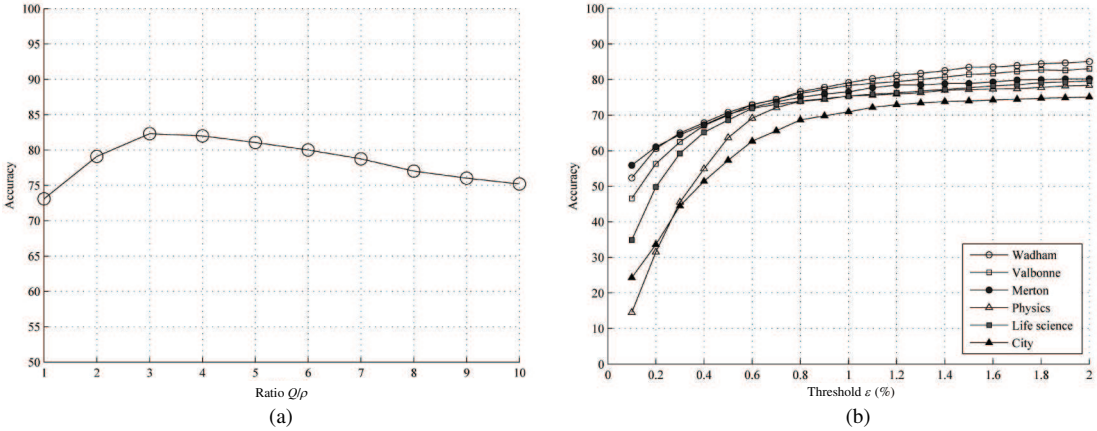
The objective of the first experiment was to demonstrate the entire process of the proposed method on the Valbonne data set. As shown in Figure 4, the structure of the scene is relatively simple (e.g., scene planes distribute mainly along two orthogonal scene directions); however, it is frequently difficult to obtain better results, including for slanted surfaces (e.g., the tower) and some details. As shown in Figure 4(b), according to the piecewise planar assumption, the superpixel consisting of pixels with homogenous color can be modeled as a spatial plane by fitting spatial points that are just projected into it. However, these fitted planes are independent of each other and usually contain a large number of redundant (nearly co-planar) ones or outliers because of computational errors or model degeneracy. Next, as shown in Figure 4(c), these planes are globally optimized using the proposed energy-based method (see Subsubsection 5.2.1) and thus consistently distributed along two orthogonal scene directions. Further, as shown in Figure 4(d), in the process of the first-stage plane inference, more planes along principal directions can be obtained, and thus, construct stronger constraints to generate a more reliable candidate plane by plane sweeping within a much reduced search range. Finally, the second-stage plane inference and the depth map refinement are reliably performed to recover the complete scene structures. Note that the texture for each plane was mapped from the corresponding superpixels. To a certain extent, visual inspection also shows that the proposed method can reconstruct a satisfactory 3D model with better accuracy.

We now present some details of plane-level depth map refinement. In Figure 5(a), the left hand parts are the close-ups of the white rectangle shown in Figure 4(c) and the right hand parts are the corresponding planes after plane-level depth map refinement. Clearly, at the stage of plane-level depth map refinement, some erroneous superpixels are further sub-segmented and the resulting sub-superpixels are assigned the correct planes. Moreover, the statistical results for the reconstruction accuracy with varying values of the ratio Q/ρ are shown in Figure 6(a). They show that a smaller Q value tends to assign the same plane to all the sub-superpixels of a superpixel, but a too large Q value could lead to some false planes for these sub-superpixels because of the smaller penalty. In our experiments, Q was set to 3ρ.

As shown in Figure 5(b) and (c), we also conducted similar experiments using the method proposed in [11], hereafter, Method A, and the method proposed in [18], hereafter Method B. Method A is a

(a) (b) (c)

Figure 5 Details of the plane-level depth map refinement and the reconstruction results using other methods. (a) The close-up of the white rectangle shown in Figure 4(c); the right hand parts are the results after plane-level depth map refinement by our method; (b) and (c) the reconstruction results produced by Method A and Method B.



(a) (b)

Figure 6 Reconstruction accuracy with different parameters and threshold values. (a) Reconstruction accuracy with different ratio Q/ρ on the Valbonne data set; (b) reconstruction accuracy with different threshold ε values on all data sets.

Table 1 Reconstruction accuracy (ε=0.02) and running time on Valbonne data set

Our method					Method A			Method B		
Planes	1st stage	2nd stage	3rd stage	Time (s)	Planes	Accuracy	Time (s)	Planes	Accuracy	Time (s)
7	0.5292	0.8228	0.8485	31.4	6	0.6223	201.8	7	0.5094	25.7

multi-view superpixel stereo method designed under the MRF framework incorporating boundary shape, photometric, and geometric constraints. In our experiments, the method performed well, since the scene has fewer principal directions pre-determined by detecting vanishing points. However, the recovered boundaries may be inaccurate because of the inaccuracy of the over-segmented superpixels. For this reason, the method fails to recover also some other details (e.g., the walls) reliably. Method B also adopts a superpixel-based MRF framework to reconstruct an approximate model of the scene. However, as discussed in Section 2, the method fails to achieve satisfactory results in comparison with that proposed in Method A and our method because of the unreliable assumption that each over-segmented superpixel is sufficiently large to contain an adequate number of spatial points for plane fitting. As shown in Figure 5(c), more superpixels (e.g., larger superpixels straddling two scene planes and some superpixels without initial spatial points) are assigned to erroneous planes at the intersection of two orthogonal scene planes. Note that, for the sake of convenient comparison, as for our method, Figure 5(b) and (c) show only the results for building regions obtained by these two methods. In fact, the unrelated regions (e.g., sky and ground) frequently greatly affect the reliability and efficiency of these methods.

Table 1 summarizes the reconstruction accuracy and the running time for the different methods. In Table 1, the reconstruction accuracy at three stages corresponds to the results of the initial noisy planes reconstruction, two-stage plane inference, and depth map refinement in our method.

Some statistical results for the reconstruction accuracy with different depth thresholds ε are shown in Figure 6(b). The results show that the reconstruction accuracy increases as ε increases, but tends to reach a plateau when ε is larger than 0.02, that is, the overall reconstruction accuracy is stable.

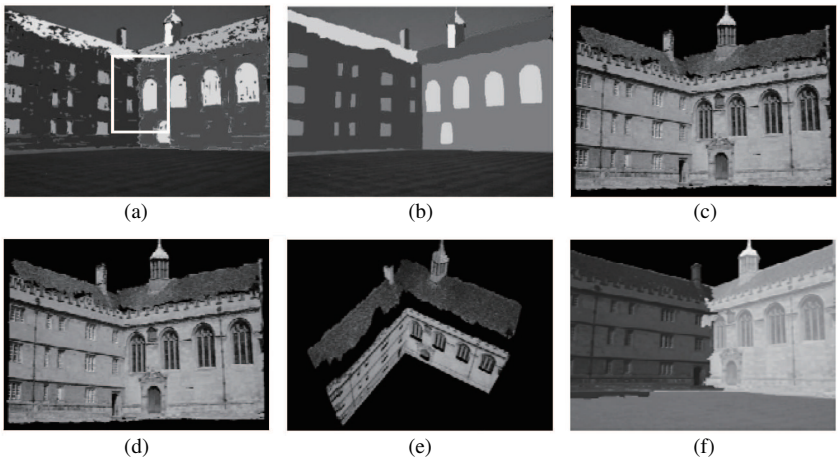


Figure 7 Reconstruction results for Wadham data set. (a) Candidate planes along principal directions; (b) shows the close-up of the white rectangle, the right hand parts are the results after plane-level depth map refinement by our method. (c) final plane reconstruction; (d) textured reconstruction; (e) top view of the textured reconstruction; (f) the reconstruction results produced by Method A.

Table 2 Quantitative comparisons (ε=0.02) on Wadham data set

Method	Planes	Accuracy	Time (s)
Ours	16	0.8251	46.2
Method A	4	0.5681	338.4

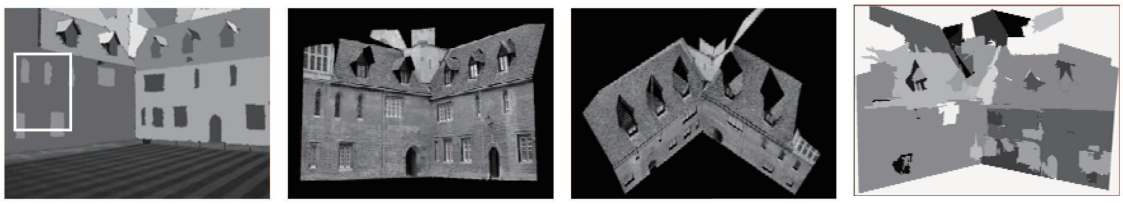


Figure 8 Reconstruction results for Merton data set. (a) Final plane reconstruction; (b) textured reconstruction; (c) top view of the textured reconstruction; (d) the reconstruction results produced by Method B.

In the second experiment, we validated the feasibility of our method on the Wadham data set used in Method A. The results are very satisfactory within expectation. In particular for some slanted surfaces (e.g., roofs) and some details (e.g., windows), our method also performed well. In fact, this scene is relatively simple, as is the Valbonne scene, but cannot be represented by only three orthogonal scene directions, and thus, Method A inevitably produces erroneous results (e.g., roofs). In addition, the recovered boundaries are still inaccurate in some regions (e.g., regions in the white rectangle shown in Figure 7(a)). As shown in Figure 7(b), these problems are resolved better by our method.

Table 2 shows the reconstruction accuracy and the running time of Method A and our method. It can be seen clearly that our method outperforms Method A in terms of both accuracy and computational efficiency.

In the third experiment, we further validated the feasibility of our method on the Merton data set used in Method B. As compared to the first two scenes, the Merton scene is slightly complex, because it contains more scene planes. As a result, our method still reliably generates better results. Here, the plane sweeping within a much reduced search range plays an important role in recovering some small planes (e.g., small scene planes in the white rectangle). In fact, as shown in Figure 8, it significantly helps to reconstruct poorly textured regions and very slanted surfaces.

In this experiment, Method B could also obtain good results, since initial spatial points distribute relatively evenly among superpixels and thus generate more reliable candidate planes. However, the

Table 3 Quantitative comparisons (ε=0.02) on Merton data set

Method	Planes	Accuracy	Time (s)
Ours	16	0.8503	42.1
Method B	24	0.6109	35.1

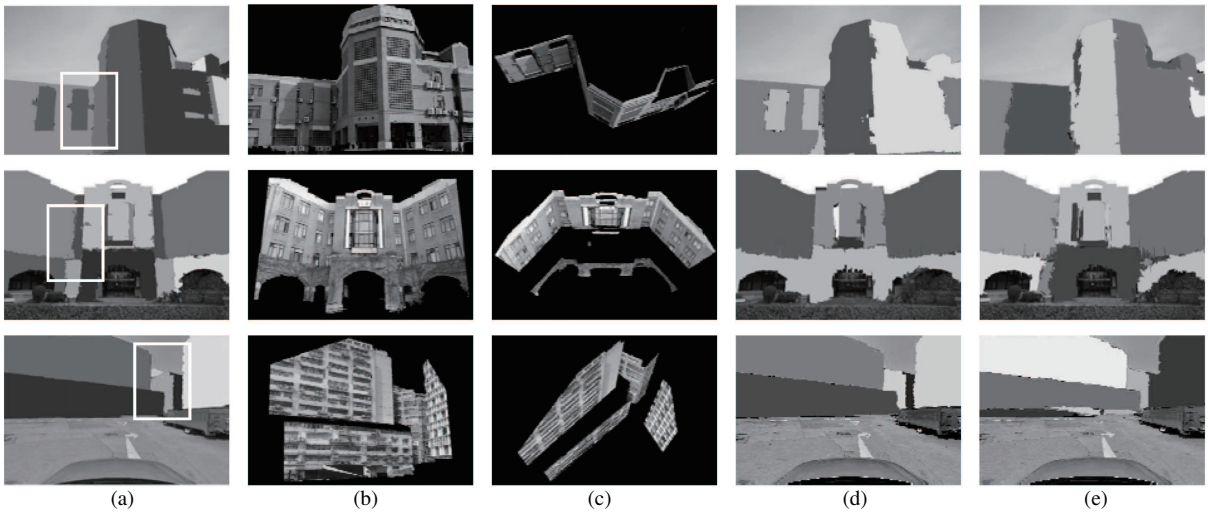


Figure 9 Reconstruction results for our own data set. (a) Final plane reconstruction; (b) textured reconstruction; (c) the top view of the textured reconstruction; (d) method A; (e) method B.

Table 4 Reconstruction accuracy (ε=0.02) and running time on our own data set

Data set	Our method			Method A			Method B		
	Planes	Accuracy	Time (s)	Planes	Accuracy	Time (s)	Planes	Accuracy	Time (s)
TPB	11	0.8513	34.1	5	0.3465	226.3	8	0.5213	29.6
LSB	12	0.8137	43.8	6	0.4933	318.8	8	0.4889	40.3
CITY	6	0.7961	57.4	4	0.6100	601.6	6	0.4122	49.5

recovered boundaries seem to be inaccurate because of the inaccuracy of the over-segmented superpixels.

Table 3 shows the reconstruction accuracy and the running time of Method B and our method. Clearly, our method is comparable with Method B, and its accuracy is higher.

The last experiment was performed using our own data sets, and all the methods were applied only in the building regions for the convenience of further comparison. These three scenes are more complex and the corresponding images also contain more negative factors, such as perspective distortion, illumination variance, poorly textured regions, and very slanted surfaces. Because of this, traditional pixel-level methods [9] frequently fail to recover the complete structures of these scenes. For the PPM methods, as shown in Figure 9 and Table 4, Method A performs well for the scenes with few orthogonal scene directions (e.g., CITY), but results in larger errors for other scenes (e.g., TPB). Conversely, Method B seems to be robust to the number of the scene directions (e.g., TPB), but fails in regions lacking spatial points, and its accuracy is lower because larger superpixels were used. In contrast, our method still performs well, in particular for some slanted surfaces and small planes (e.g., the wall in the white rectangle), and the reconstructed results are also satisfactory. To some extent, the existing problems in Method A and Method B could be reliably solved by virtue of the robust two-stage plane optimization and inference. Moreover, in Table 4, we can see that our method is a little slower, but has higher accuracy than Method B.

Figure 10 also shows the results corresponding to multiple images. It can be seen clearly that the structures of scenes can be recovered completely. The results of the experiment demonstrate that a satisfactory scene model can be obtained by integrating each result corresponding to a single image with better accuracy.

(a) (b) (c)

Figure 10 Reconstruction results of three neighboring images. (a) TPB; (b) LSB; (c) CITY.

Table 5 Numerical results

Data set	Initialization			Planes			Time		
	PTS	SP	PSP (%)	INI	PP	FIN	PI1 (s)	PI2 (s)	PR (s)
Valbonne	405	1385	61.1	1002	5	7	14.3	12.8	0.66
Wadham	5048	3460	69.8	2919	12	16	22.4	19.2	1.65
Merton	4982	6109	62.5	5091	14	20	25.7	21.0	2.03
LSB	1669	4254	46.2	3321	8	12	18.2	13.9	0.91
TPB	1150	3719	38.9	2673	6	11	27.6	14.4	1.45
CITY	3622	4511	51.0	2923	5	6	34.3	19.1	1.98

Table 5 shows some numerical results. Here, PTS stands for the number of initial spatial points, SP and PSP respectively stand for the number of superpixels and the percentage of superpixels with initial spatial points. INI and PP respectively stand for the number of initial planes and the number of principal scene planes obtained by energy-based plane optimization, and FIN is the number of planes after the two-state plane inference and plane-level depth map refinement. We provide the approximate running time costs for the two-stage (PI1 and PI2) plane inference and plane-level refinement (PR) in seconds for piecewise planar depth map construction of the current image.

In Table 5, we can see that the proposed method is generally stable and its performance is better because of the effective multi-level superpixel-based energy formulation.

In summary, as compared to the state-of-the-art piecewise planar stereo methods, our proposed method can efficiently and effectively recover more complete structures of a scene containing poorly textured regions and very slanted surfaces, and is more robust to complex scenes with more negative factors (e.g., illumination variance and various scene structures)

8 Conclusion

We proposed an automatic multi-view piecewise planar stereo method under an energy minimization framework for the complete reconstruction of urban scenes. Our method can effectively handle the many difficulties (e.g., textureless regions and slanted surfaces) encountered in standard pixel-level stereo methods. In addition, unlike other piecewise planar stereo methods that recover only a simplistic 3D model, our method generates a better result and is more robust to complex scenes because it incorporates more scene priors and performs a robust two-stage plane inference and an effective plane-level depth map refinement. The limitations of our method include the following. (1) The performance is influenced by the number of initial spatial points. In general, spatial points that are too sparse frequently fail to construct stronger constraints for the plane inference and optimization. (2) The reconstruction accuracy is limited by the nature of the PPM assumption. In practice, the first limitation could be overcome by densifying the initial spatial points using point propagation methods. To resolve the second limitation, a pixel-level based post-processing step could be used to enrich our reconstructed piecewise planar scene. Finally, by incorporating more semantic priors into the energy optimization framework, the accuracy of the plane inference and refinement could also be increased.

Acknowledgements This work was supported by National Natural Science Foundation of China (Grant Nos. 61421004, 61333015, 61273280, 61103143, U1404620, U1404622), Development Project of Henan Provincial Department of Science and Technology (Grant No. 152102310381), and Scientific Research Starting Foundation for Advanced Talents of Zhoukou Normal University (Grant No. zknuc2015103).

Conflict of interest The authors declare that they have no conflict of interest.

References

1 Hong L, Chen G. Segment-based stereo matching using graph cuts. In: Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition, Washington, 2004. 74–81
2 Zitnick C L, Kang S B. Stereo for image-based rendering using image over-segmentation. Int J Comput Vision, 2007, 75: 49–65
3 Klaus A, Sormann M, Karner K. Segment-based stereo matching using belief propagation and a self-adapting dissimilarity measure. In: Proceedings of the 18th International Conference on Pattern Recognition (ICPR'06), Hong Kong, 2006. 15–18
4 Wang Z F, Zheng Z G. A region based stereo matching algorithm using cooperative optimization. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, Anchorage, 2008. 1–8
5 Boykov Y, Veksler O, Zabih R. Fast approximate energy minimization via graph cuts. IEEE Trans Pattern Anal Mach Intell, 2001, 23: 1222–1239
6 Bleyer M, Rhemann C, Rother C. PatchMatch stereo - stereo matching with slanted support windows. In: Proceedings of British Machine Vision Conference, Dundee, 2011. 1–11
7 Çiğla C, Zabulis X, Alatan A A. Segment-based stereo-matching via plane and angle sweeping. In: Proceedings of 3DTV Conference, Kos Island, 2007. 1–4
8 Furukawa Y, Curless B, Seitz S, et al. Manhattan-world stereo. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, Miami, 2009. 1422–1429
9 Furukawa Y, Ponce J. Accurate, dense, and robust multi-view stereopsis. IEEE Trans Pattern Anal Mach Intell, 2009, 32: 1362–1376
10 Mičušík B, Košecká J. Piecewise planar city 3D modeling from street view panoramic sequences. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, Miami, 2009. 2906–2912
11 Mičušík B, Košecká J. Multi-view superpixel stereo in urban environments. Int J Comput Vision, 2010, 89: 106–119
12 Gallup D, Frahm J M, Mordohai P, et al. Real-time plane-sweeping stereo with multiple sweeping directions. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, Minneapolis, 2007. 1–8
13 Gallup D, Frahm J M, Pollefeys M. Piecewise planar and non-planar stereo for urban scene reconstruction. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, San Francisco, 2010. 1418–1425
14 Sinha S N, Steedly D, Szeliski R. Piecewise planar stereo for image-based rendering. In: Proceedings of IEEE 12th International Conference on Computer Vision, Kyoto, 2009. 1881–1888
15 Kim H, Xiao H, Max N. Piecewise planar scene reconstruction and optimization for multi-view stereo. In: Proceedings of the 11th Asian Conference on Computer Vision. Berlin: Springer, 2012. 191–204
16 Chauve A L, Labatut P, Pons J P. Robust piecewise-planar 3D reconstruction and completion from large-scale unstructured point data. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR), San Francisco, 2010. 1261–1268
17 Kowdle A, Sinha S N, Szeliski R. Multiple view object cosegmentation using appearance and stereo cues. In: Proceedings of the 12th European Conference on Computer Vision. Berlin: Springer, 2012. 789–803
18 Bodis-Szomoru A, Riemenschneider H, van Gool L. Fast, approximate piecewise-planar modeling based on sparse structure-from-motion and superpixel. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, Columbus, 2014. 469–476
19 Gould S, Fulton R, Koller D. Decomposing a scene into geometric and semantically consistent regions. In: Proceedings of IEEE 12th International Conference on Computer Vision, Kyoto, 2009. 1–8
20 Comaniciu D, Meer P. Mean shift: a robust approach toward feature space analysis. IEEE Trans Pattern Anal Mach Intell, 2002, 24: 603–619
21 Fischler M A, Bolles R C. Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. Commun ACM, 1981, 24: 381–395
22 Isack H, Boykov Y. Energy-based geometric multi-model fitting. Int J Comput Vision, 2010, 92: 123–147
23 Delong A, Osokin A, Isack H N, et al. Fast approximate energy minimization with label costs. Int J Comput Vision, 2012, 96: 1–27
24 Pham T T, Chin T J, Yu J, et al. The random cluster model for robust geometric fitting. IEEE Trans Pattern Anal Mach Intell, 2014, 36: 1658–1671
25 Tola E, Strecha C, Fua P. Efficient large-scale multi-view stereo for ultra high-resolution image sets. Mach Vision Appl, 2012, 23: 903–920
26 Frey B J, Dueck D. Clustering by passing messages between data points. Science, 2007, 315: 972–976
27 Kohli P, Ladicky L, Torr P H S. Robust higher order potentials for enforcing label consistency. Int J Comput Vision, 2009, 82: 302–324