Manhattan Scene Understanding Via XSlit Imaging

Jinwei Ye Yu Ji Jingyi Yu

University of Delaware, Newark, DE 19716, USA

{jye,yuji,yu}@cis.udel.edu

Abstract

A Manhattan World (MW) [3] is composed of planar surfaces and parallel lines aligned with three mutually orthogonal principal axes. Traditional MW understanding algorithms rely on geometry priors such as the vanishing points and reference (ground) planes for grouping coplanar structures. In this paper, we present a novel single-image MW reconstruction algorithm from the perspective of non-pinhole cameras. We show that by acquiring the MW using an XSlit camera, we can instantly resolve coplanarity ambiguities. Specifically, we prove that parallel 3D lines map to 2D curves in an XSlit image and they converge at an XSlit Vanishing Point (XVP). In addition, if the lines are coplanar, their curved images will intersect at a second common pixel that we call Coplanar Common Point (CCP). CCP is a unique image feature in XSlit cameras that does not exist in pinholes. We present a comprehensive theory to analyze XVPs and CCPs in a MW scene and study how to recover 3D geometry in a complex MW scene from XVPs and CCPs. Finally, we build a prototype XSlit camera by using two layers of cylindrical lenses. Experimental results on both synthetic and real data show that our new XSlit-camera-based solution provides an effective and reliable solution for MW understanding.

1. Introduction

A pinhole camera collects rays passing through a common *Center-of-Projection* (CoP) and has been the dominating imaging model for computer vision tasks. The pinhole model is popular for two main reasons. First, pinhole geometry is simple; it is uniquely defined by only 3 parameters (the position of CoP in 3D) and its imaging process can be uniformly described by the classic 3×4 pinhole camera matrix [12]. Second, human eyes act as a virtual pinhole camera, *e.g.*, they observe lines as lines and parallel lines converging at a vanishing point. Pinhole cameras hence are also referred to as perspective cameras.

The pinhole imaging model, however, is rare in insect eyes. Compound eyes, which may consist of thousands of individual photoreceptor units or ommatidia are much more common. Images perceived are a combination of inputs from the numerous individual "eye units" pointing towards different directions. These multi-perspective imaging models provide unique advantages for perceiving and interpreting scene geometry [28]. In this paper, we demonstrate using a special multi-perspective camera, the XSlit camera [30], for reconstructing the Manhattan World scenes.

The Manhattan World. A Manhattan World (MW) [3] is composed of planar surfaces and parallel lines aligned with three mutually orthogonal principal axes. The MW model fits well to many man-made (interior/exterior) environments that exhibit strong geometry regularity such as flat walls, axis-aligned windows and sharp corners. Tremendous efforts have been focused on reconstructing MW from images [1, 5, 10, 11] and using the MW assumption for camera calibration [3, 15, 24]. The main challenge is that MW generally exhibits repeated line patterns but lacks textures for distinguishing between them, making it difficult to directly apply stereo matching. Furukawa *et al*. [10, 11] assign a plane to each pixel and then apply graph-cut on discretized plane parameters.

MW reconstruction/understanding from a single image is even more challenging. Most previous approaches exploit monocular cues such as the vanishing points and the reference planes (*e.g*. the ground) for approximating scene geometry. Hoime *et al*. [13, 14] use image attributes (color, edge orientation, *etc*.) to label image regions with different geometric classes (sky, ground, and vertical) and then "pop-up" the vertical regions to generate visually pleasing 3D reconstructions. Delage *et al*. [5, 6] extend this technique to indoor scenes. Kosecka and Zhang [15] detect line structures in the image for recovering the vanishing points and camera parameters [29]. Criminisi *et al*. [4] use the vanishing points and ground plane as priors for recovering affine scene structures. Saxena *et al*. [22, 23] apply machine learning techniques to infer depths from image features and

use the Markov Random Field (MRF) to determine the location and orientation of planar regions. Lee *et al*. [16] and Flint *et al*. [9] search for the most feasible combination of line segments for indoor MW understanding.

Our Approach. We present a novel single-image MW reconstruction algorithm from the perspective of non-pinhole imaging. We observe that the core challenge in pinhole-based solutions is coplanar ambiguities: although one can easily detect the vanishing point of a group of parallel 3D lines, there is an ambiguity on which lines belong to the same plane. We show that this ambiguity can be naturally resolved if we use a multi-perspective camera to acquire the scene.

Conceptually, 3D parallel lines will be mapped to 2D curves in a multi-perspective camera and these curves will intersect at multiple points instead of a single vanishing point. For example, Caglioti *et al*. [2] examine the curves in a non-centric catadioptric camera for line localization. Swaminathan *et al*. [26] develop a solution for axial non-centric cameras. In this paper, we show how to group parallel 3D lines on the same plane by analyzing their images in a special multi-perspective cameras, the XSlit camera [30]. We show that same as in the pinhole camera, images of parallel lines in an XSlit image, although curved, will still converge at a vanishing point, *i.e*., the XSlit Vanishing Point or XVP. What is different though is that images of coplanar 3D lines will generally intersect at a *second* common point that we call Coplanar Common Point or CCP. CCP is a special feature of XSlits that does not exist in pinholes. We then present a comprehensive theory to analyze XVPs and CCPs of a MW scene. We show that the geometry of 3D lines can be directly recovered from their XVPs and CCPs. We further develop a robust algorithm to distinguish the two types of points of complex MW scenes using a single XSlit image. Finally, we construct a prototype XSlit camera by using two layers of cylindrical lenses. Experimental methods on both synthetic and real data show that our XSlit camera based solution provides an effective and reliable solution for MW scene understanding.

Our contributions include:

- A new theory to characterize the XVP and CCP of coplanar parallel 3D lines in an XSlit image.
- A class of robust techniques for locating, separating, and finally utilizing the XVPs and CCPs for MW scene reconstruction from a single image.
- A prototype XSlit camera for validating our theory.

2. XSlit Imaging

Traditional approaches use the projection from 3D points to 2D pixels to model the imaging process in a camera. In this paper we decompose the projection process into two components: the mapping from a 3D point to a ray collected in the camera and the mapping from the ray to a pixel. We use the two-plane parametrization (2PP) [17] for parameterizing rays. In 2PP, each ray is parameterized as $[u, v, s, t]$, where $[u, v]$ and $[s, t]$ are the intersections with the two parallel planes Π_{uv} and Π_{st} lying at $z = 0$ and $z = 1$ respectively. $[u, v, s, t]$ can be viewed as a two-point representation of a line. To further simplify our analysis, we use $[u, v, \sigma, \tau]$ parametrization where $\sigma = s - u$ and $\tau = t - v$. We choose Π_{uv} as the default image (sensor) plane so that $[\sigma, \tau, 1]$ can be viewed as the direction of the ray.

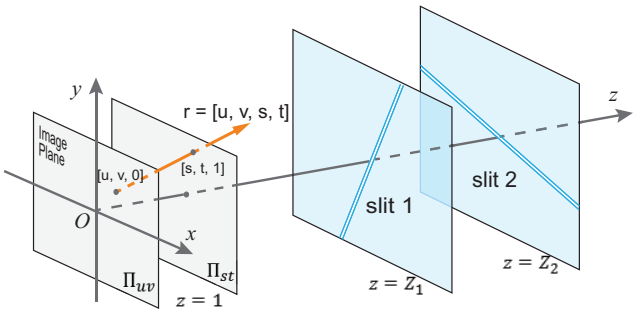

Figure 1. An XSlit Camera collects rays simultaneously passing through two oblique slits. We use a special two-plane parametrization (2PP) to represent all XSlit rays.

2.1. XSlit Camera Geometry

An XSlit camera collects rays that simultaneously pass through two oblique (neither parallel nor coplanar) slits in 3D space [30]. Given two slits l_1 and l_2, we construct the 2PP as follows: we choose Π_{uv} and Π_{st} that are parallel to both slits but do not contain them, as shown in Fig. 1. Next, we orthogonally project both slits on Π_{uv} and use their intersection point as the origin of the coordinate system. We assume l_1 and l_2 lie at $z = Z_1$ and $z = Z_2$ with directions $[d_x^1, d_y^1, 0]$ and $[d_x^2, d_y^2, 0]$, where $Z_1 \neq Z_2$ and $d_x^1 d_y^2 - d_y^1 d_x^2 \neq 0$.

XSlit Ray Constraints. We first explore ray geometry constraints for all rays in an XSlit camera. Since each ray $[u, v, \sigma, \tau]$ simultaneously passes through l_1 and l_2, there must exist some λ_1 and λ_2 so that

$$\begin{cases} u + Z_1\sigma = \lambda_1 d_x^1; & v + Z_1\tau = \lambda_1 d_y^1 \\ u + Z_2\sigma = \lambda_2 d_x^2; & v + Z_2\tau = \lambda_2 d_y^2 \end{cases} \tag{1}$$

Eliminating λ_1 and λ_2, we obtain two linear constraints in $[u, v, \sigma, \tau]$ as

$$\begin{cases} \sigma &= (Au + Bv)/E \\ \tau &= (Cu + Dv)/E \end{cases} \tag{2}$$

where

$$A = d_x^2 d_y^1 Z_2 - d_x^1 d_y^2 Z_1, \quad B = d_x^1 d_x^2 (Z_1 - Z_2),$$
$$D = d_x^2 d_y^1 Z_1 - d_x^1 d_y^2 Z_2, \quad C = d_y^1 d_y^2 (Z_2 - Z_1),$$
$$E = (d_x^1 d_y^2 - d_x^2 d_y^1) Z_1 Z_2$$

Recall that the two slits are oblique, therefore $E \neq 0$. We call Eqn. (2) the *XSlit Ray Constraints* (XSRC). XSRC has been introduced in various forms in previous studies, *e.g.*, as projection model in [30], as general linear constraints in [27], and as ray regulus in [21].

2.2. Other Ray Geometry Constraints

Rays Passing Through a 3D Line. The focus of this paper is to study the projection of 3D lines. We therefore derive the constraint for rays to pass through a line and then combine it with XSRC to study its image. Given a 3D line l [1], we consider two cases.

Case 1: If $l \parallel \Pi_{uv}$, we can represent it as $l : (x_l, y_l, z_l) + \lambda(d_x^l, d_y^l, 0)$. If a ray $r[u, v, \sigma, \tau]$ passes through l, there must exist some λ and λ_l so that

$$[u, v, 0] + \lambda[\sigma, \tau, 1] = [x_l, y_l, z_l] + \lambda_l [d_x^l, d_y^l, 0] \tag{3}$$

It is easy to see that $\lambda = z_l$. Eliminating λ_l, we obtain a linear constraint:

$$\frac{u}{d_x^l} - \frac{v}{d_y^l} + \frac{z_l \sigma}{d_x^l} - \frac{z_l \tau}{d_y^l} - \frac{x_l}{d_x^l} + \frac{y_l}{d_y^l} = 0 \tag{4}$$

We call Eqn.(4) the *Parallel Line Constraint*. We can further combine it with XSRC (Eqn. (2)) to find all rays passing through l. These three linear constraints yield to a linear constraint in u and v, *i.e.*, the image of the line. Therefore the image of $l \parallel \Pi_{uv}$ is still a line. Further, its slope depends only on the direction of l and therefore all 3D lines parallel to Π_{uv} will be mapped to 2D parallel lines.

Case 2: If $l \nparallel \Pi_{uv}$, we can parameterize it under 2PP as $[u_l, v_l, \sigma_l, \tau_l]$. Similar to case 1, there must exist some λ and λ_l so that

$$[u, v, 0] + \lambda[\sigma, \tau, 1] = [u_l, v_l, 0] + \lambda_l [\sigma_l, \tau_l, 1] \tag{5}$$

We have $\lambda = \lambda_l$ and eliminating λ andλ_l, we obtain a bilinear constraint:

$$\frac{u - u_l}{v - v_l} = \frac{\sigma - \sigma_l}{\tau - \tau_l} \tag{6}$$

We call Eqn. (6) the *Non-Parallel Line Constraint*. Using XSRC (Eqn. (2)), we can solve for σ and τ with u and v and substitute them into Eqn. (6). We then obtain a conic curve in u and v, *i.e.*, the image of l as

$$\begin{aligned} &Cu^2 + (D - A)uv - Bv^2 + (Av_l - Cu_l - E\tau_l)u \\ &\quad + (Bv_l - Du_l + E\sigma_l)v + E(u_l\tau_l - v_l\sigma_l) = 0 \end{aligned} \tag{7}$$

To determine the type of the conic, we compute

$$J = (D - A)^2 - 4BC = (d_x^1 d_y^2 - d_x^2 d_y^1)^2 (Z_1 - Z_2)^2 > 0 \tag{8}$$

Therefore, the conic can only be hyperbolas. Eqn. (7) further reveals that the quadratic coefficients of the hyperbolas only depend on the XSlit intrinsic parameters (A, B, C, and D), *i.e.*, they are identical for all 3D lines. In Sec. 3.3, we use this property for fitting hyperbolas from the acquired XSlit images. Notice though that we cannot directly reconstruct a 3D line from its hyperbola image: a line has four unknowns u_l,v_l,σ_l, and τ_l while we only have three equations (the u and v coefficients and the constant term in Eqn. (7)). Similar ambiguity also exists in pinhole cameras.

Rays Lying on A Plane. The last crucial ray geometry constraint is for rays lying on a common plane. This lies at the core of this paper as our goal is to disambiguate parallel 3D lines lying on different planes. Given a plane $\Pi \nparallel \Pi_{uv}$: $n_x x + n_y y + n_z z + d = 0$, with $\vec{n} = [n_x, n_y, n_z]$ being the normal and d the offset, we can intersect Π with Π_{uv} at line:

$$n_x u + n_y v + d = 0 \tag{9}$$

All rays $[u, v, \sigma, \tau]$ that lie on Π must originate from this line, *i.e.*, they have to satisfy Eqn. (9). Further, the direction of rays must be orthogonal to $\vec{n}$. Therefore, we have

$$n_x \sigma + n_y \tau + n_z = 0 \tag{10}$$

We call these two linear constraints the *Rays-on-Plane Constraints*.

2.3. XSlit Vanishing Points (XVP)

Next, we use the ray geometry constraints for studying the vanishing points of parallel 3D lines in an XSlit image.

Theorem 1. *Given a set of parallel lines $\mathcal{L} \nparallel \Pi_{uv}$, their images on Π_{uv} have a vanishing point.*

Proof. Assume $\mathcal{L}$ have a common direction $[\sigma^*, \tau^*, 1]$ but different origins $[u_l, v_l, 0]$, all points on each line can be written as

$$P(x, y, z) = [u_l, v_l, 0] + \lambda[\sigma^*, \tau^*, 1] \tag{11}$$

Using the XSlit point projection equation, we have a line image as

$$\begin{cases} u &= \dfrac{(D\sigma^* - B\tau^*)E\lambda^2 + (Du_l - Bv_l + E)E\lambda + E^2 u_l}{(AD - BC)\lambda^2 + (A + D)E\lambda + E^2} \\ v &= \dfrac{(A\tau^* - C\sigma^*)E\lambda^2 + (Av_l - Cu_l + E)E\lambda + E^2 v_l}{(AD - BC)\lambda^2 + (A + D)E\lambda + E^2} \end{cases} \tag{12}$$

[1] Although slits are essentially lines, we distinguish them two for clarity: slits refer to the XSlit camera geometry and lines refer to 3D scene.

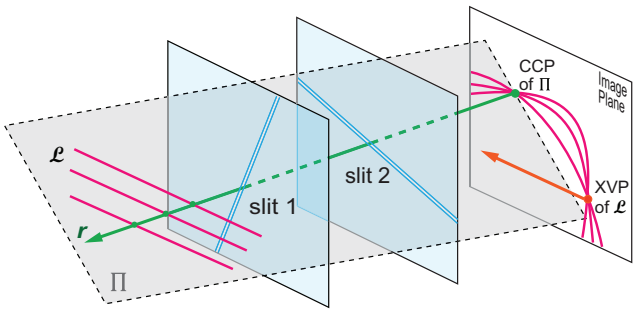

Figure 2. In an XSlit camera, the images of coplanar parallel lines appear curved and will intersect at two common points, the XSlit Vanishing Point (XVP) and the Coplanar Common Point (CCP).

Since both the denominators and the numerators are quadratic in λ, when $\lambda \to \infty$, we have:

$$\begin{cases} u_\infty &= \dfrac{E(D\sigma^* - B\tau^*)}{AD - BC} \\ v_\infty &= \dfrac{E(A\tau^* - C\sigma^*)}{AD - BC} \end{cases} \tag{13}$$

where $AD - BC = Z_1 Z_2 (d_{1_x} d_{1_2} - d_{1_y} d_{2_x})^2 \neq 0$ as the two slits are oblique. The results are independent of the origin of the line and therefore the images of parallel 3D lines, although being hyperbolas, have a vanishing point. It corresponds to the intersection point at infinite of these lines. □

2.4. Coplanar Common Points (CCP)

What differs XSlit from pinhole cameras and hence makes it appealing is that parallel 3D lines lying on a plane will converge at a *second common point* in an XSlit image.

Theorem 2. *Given a set of lines $\mathcal{L}$ that lie on plane Π unparallel to the two slits, their images in the XSlit camera generally intersect at a second common point, the Coplanar Common Point or CCP.*

Proof. Notice that the CCP corresponds to some ray r that 1) is collected by the XSlit, 2) lies on Π, and 3) will intersect all lines in $\mathcal{L}$, as shown in Fig. 2. For 1) and 2), we can combine the XSRC (Eqn. (2)) and the Rays-on-Plane Constraints (Eqn. (9) and (10)) and solve for the ray. Notice that both sets of constraints are linear in u, v, σ, τ, therefore, we form a linear system:

$$\mathbf{A\Phi} = \mathbf{b} \tag{14}$$

where

$$\mathbf{A} = \begin{bmatrix} A & B & -E & 0 \\ C & D & 0 & -E \\ 0 & 0 & n_x & n_y \\ n_x & n_y & 0 & 0 \end{bmatrix}, \quad \mathbf{\Phi} = \begin{bmatrix} u \\ v \\ \sigma \\ \tau \end{bmatrix}, \quad \mathbf{b} = \begin{bmatrix} 0 \\ 0 \\ -n_z \\ -d \end{bmatrix}$$

Computing the determinant of $\mathbf{A}$, we have

$$\begin{aligned} \det(\mathbf{A}) &= n_x(Bn_x + Dn_y) - n_y(An_x + Cn_y) \\ &= (Z_2 - Z_1)(n_x d_x^1 + n_y d_y^1)(n_x d_x^2 + n_y d_y^2) \end{aligned} \tag{15}$$

For any plane Π that is not parallel to either slit, $\det(\mathbf{A}) \neq 0$, thus we obtain a unique r from the XSlit and lies on Π. CCP is then the intersection of r with Π_{uv}:

$$\begin{cases} u &= \dfrac{En_y n_z - d(Bn_x + Dn_y)}{n_x(Bn_x + Dn_y) - n_y(An_x + Cn_y)} \\ v &= -\dfrac{En_x n_z - d(An_x + Cn_y)}{n_x(Bn_x + Dn_y) - n_y(An_x + Cn_y)} \end{cases} \tag{16}$$

r generally intersects $\mathcal{L}$ except for the singular case that $\mathcal{L} \parallel r$. □

CCP is a unique image feature in XSlit cameras that does not exist in pinholes. Notice that pinhole can be viewed as a *degenerate* XSlit camera, *i.e.*, when the two slits intersect ($Z_1 = Z_2$). In this case, $\det(\mathbf{A}) = 0$ by Eqn. (15) and therefore CCP does not exist. We can also interpret the result from the view of geometry. For the pinhole camera to have a ray to lie completely on a 3D plane, the plane has to pass the pinhole. In contrast, for a plane to have CCP in an XSlit camera, it only needs to be unparallel to the two slits, a condition that generally holds.

3. Manhattan World Reconstruction

Next, we use CCPs and XVPs for MW reconstruction.

3.1. Recovering Planes

We first show how to recover a plane Π that contains parallel 3D lines $\mathcal{L}$ using their CCP and XVP.

Theorem 3. *Given a set of coplanar parallel lines $\mathcal{L}$, if they have a CCP, it will not coincide with their XVP.*

Proof. Notice that the CCP corresponds to a ray that intersect all lines at a finite distance whereas the XVP corresponds to a ray that intersects the lines at the infinite distance. Therefore the two points correspond to two different rays and thus will not coincide. □

Next, we show how to recover the plane Π. We first compute the normal of Π. Given the XVP $[u_v, v_v]$ and the XSlit intrinsic parameters (A, B, C, D, and E), by Eqn. (13), we can directly compute the direction of $\mathcal{L}$ $\vec{l}_v = [\sigma_v, \tau_v, 1]$ as

$$\begin{cases} \sigma_v &= (Au_v + Bv_v)/E \\ \tau_v &= (Cu_v + Dv_v)/E \end{cases} \tag{17}$$

Now consider the CCP $[u_c, v_c]$ that also corresponds to a ray lying on Π. Therefore, we can compute its direction $\vec{l}_c = [\sigma_c, \tau_c, 1]$ by using the XSRC (Eqn. (2)) as

$$\begin{cases} \sigma_c &= (Au_c + Bv_c)/E \\ \tau_c &= (Cu_c + Dv_c)/E \end{cases} \tag{18}$$

By Theorem 3, the XVP and CCP will not coincide and therefore $\vec{l}_v$ and $\vec{l}_c$ will not be collinear. The normal of Π is thus $\vec{n} = \vec{l}_v \times \vec{l}_c$. Finally, since the CCP lies on Π, we can compute Π's offset using its coordinate as

$$d = n_x u_c + n_y v_c \tag{19}$$

3.2. Manhattan World

An important requirement for applying the plane recovery algorithm is to know which point is CCP and which one is XVP, as they both appear as the common intersection points of the curves. In particular, if only one set of coplanar parallel lines is available, we cannot distinguish CCP from XVP. In fact, exchanging the assignment will result in different but both valid planes. In reality, a typical Manhattan scene contains multiple sets of coplanar parallel lines for resolving this ambiguity. Specifically, we exploit structural regularity as follows.

Manhattan World (MW) Assumption [3]: We assume that objects in the scene are composed of planes and lines aligned with three mutually orthogonal principal axes, *i.e.*, L_1, L_2, and L_3 of $[\sigma_i, \tau_i, 1]$, $i = 1, 2, 3$ respectively, and

$$\sigma_i \sigma_j + \tau_i \tau_j + 1 = 0, \text{ where } i,\ j = 1, 2, 3 \text{ and } i \neq j. \tag{20}$$

The three axes will map to three XVPs, namely, $\text{XVP}_1[u_1, v_1]$, $\text{XVP}_2[u_2, v_2]$, and $\text{XVP}_3[u_3, v_3]$.

Theorem 4. *The CCP_{L_1} of a plane with normal L_1,* i.e.*, it is parallel to $\Pi_{L_2L_3}$, lies on the line XVP$_2$-XVP$_3$.*

Proof. We prove the theorem by showing that the three points $\text{CCP}_{L_1}[u_{c_1}, v_{c_1}]$, XVP_2, and XVP_3 are co-linear. We first compute

$$\det(\mathbf{M}) = \begin{vmatrix} u_{c_1} & u_2 & u_3 \\ v_{c_1} & v_2 & v_3 \\ 1 & 1 & 1 \end{vmatrix} \tag{21}$$

Since $\Pi_{L_2L_3}$ has normal $[\sigma_1, \tau_1, 1]$, by CCP Eqn. (16), we have $[u_{c_1}, v_{c_1}]$ as

$$\begin{cases} u_{c_1} &= \dfrac{E\tau_1 - d(B\sigma_1 + D\tau_1)}{\sigma_1(B\sigma_1 + D\tau_1) - \tau_1(A\sigma_1 + C\tau_1)} \\ v_{c_1} &= -\dfrac{E\sigma_1 - d(A\sigma_1 + C\tau_1)}{\sigma_1(B\sigma_1 + D\tau_1) - \tau_1(A\sigma_1 + C\tau_1)} \end{cases} \tag{22}$$

Similarly, we can rewrite u_2, v_2, u_3, and v_3 in terms of σ_2, τ_2, σ_3, and τ_3 using XVP Eqn. (13) and substitute them into Eqn. (21). Reusing the orthogonality condition (Eqn. (20)), we derive that $\det(\mathbf{M}) = 0$ □

Notice that Theorem 4 is independent of the offset d of the plane. Therefore, the CCPs for all planes parallel to $\Pi_{L_2L_3}$ will lie on the line XVP$_2$-XVP$_3$. Similar conclusions hold for planes parallel to the other two principal planes $\Pi_{L_1L_2}$ and $\Pi_{L_3L_1}$. Therefore, Theorem 4 provides an effective and robust means for disambiguating CCPs and XVPs: in a MW scene, all CCPs and XVPs should lie on a triangle where XVPs correspond to the triangle vertices and CCPs lie on triangle edges (or the extension of edges).

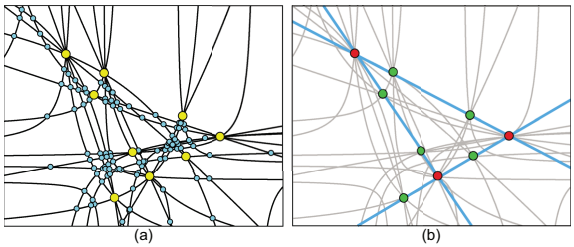

Figure 3. Locating XVPs and CCPs. (a) shows the intersection points between a group of hyperbolas. The blue dots are the outliers and the yellow ones are either XVPs or CCPs; (b) We fit three lines using only the yellow dots. The resulting triangle vertices (red) correspond to XVPs and the edge points to CCPs (green).

3.3. MW Scene Reconstruction

In order to use Theorem 4 for reconstructing a MW scene, we strategically tilt our XSlit camera to make the slits unparallel to the principal axes L_1, L_2, and L_3 of the buildings so that we can use XVPs and CCPs of different building faces.

Conic Fitting. We first fit conics to images of the lines and compute pairwise intersections. We have shown in Sec. 2.2 that the images of lines are hyperbolas with the form: $\tilde{A}u^2 + \tilde{B}uv + \tilde{C}v^2 + \tilde{D}u + \tilde{E}v + \tilde{F} = 0$ where $\tilde{A}$, $\tilde{B}$, and $\tilde{C}$ are uniquely determined by the XSlit camera intrinsics that can be precomputed and are identical for all hyperbolas. We apply a similar curve fitting scheme as [7] by forming an over-determined linear system of conic coefficients using the sampled points on the curves. We then apply the Singular Value Decomposition (SVD) to solve for conic parameters.

XVP/CCP Identification. Once we fit all conics, we compute their intersection points and locate the XVPs and CCPs. Notice that in addition to XVPs and CCPs, every two conics that correspond to two unparallel 3D lines may also intersect. Since their intersection point will not be shared by other conics, we can remove the intersections that only appear once to eliminate outliers.

By Theorem 4, all CCPs are located on the edges of the triangle determined by the three XVPs. Therefore, we fit three lines using the rest intersections and use the resulting triangle vertices and edges to separate the XVPs from the CCPs. Fig. 3 illustrates this process for a simple scene composed of 18 lines on 6 planes. Each plane has 3 parallel lines lying on it and the directions of all lines are aligned with the three principal axes.

Plane Reconstruction. To reconstruct a MW scene from

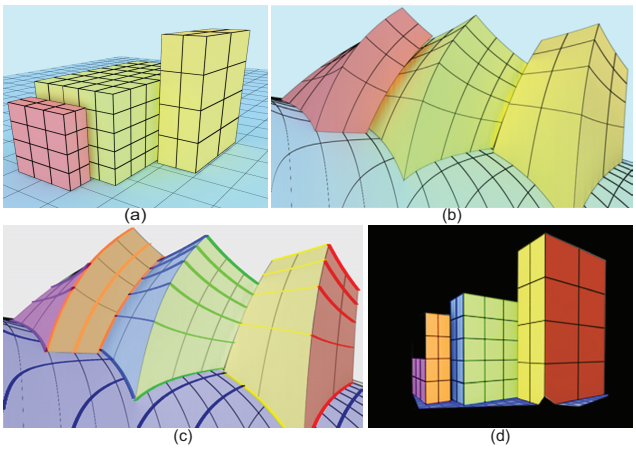

Figure 4. Manhattan scene reconstruction from a single XSlit image. (a) A pinhole image of the scene. (b) An XSlit image. (c) Automatically detected planes using our approach are highlighted in different colors. (d) We re-render the recovered model using a perspective camera.

a single XSlit image, we directly map each CCP back to a plane. Theorem 2 reveals that every CCP corresponds to a unique 3D plane in the scene. Specifically, for each detected CCP, we can combine it with one of XVPs (triangle vertices) for computing the plane equation as shown in Sec. 3.1. We further map each curve segment back to a 3D line segment by intersecting the XSlit rays originated from the conic with the reconstructed plane. The endpoints of the line segments can also be used for truncating the recovered planes.

4. Experiments

We have validated our approach on both synthetic and real data.

4.1. Synthetic XSlit Images

We start with testing our method on a simple MW scene. We place three axis-aligned boxes in the scene and each face of the box contains multiple grid lines. To simulate an XSlit image, we modify the POV-Ray raytracer (www.povray.org) by adding an XSlit camera model. Fig. 4(b) shows a sample ray-traced image using the XSlit camera. In this example, we use two orthogonal slits (*i.e.*, a POX-Slit [30]) and the two slits and the image plane are evenly spaced. We rotate the camera 45° around the y axis and 15° around both x and z axes so that axis-aligned lines will have XVPs and CCPs in the XSlit image.

We use the Canny edge detector to locate the curve. For each curve, we use multiple pixels that lie on it and apply the curve fitting technique (Sec. 3.3) for obtaining its parameters. Next, we compute pairwise curve intersections and apply the XVP/CCP detection algorithm (Sec. 3.3) to find the CCPs. Each detected CCP is then mapped back to

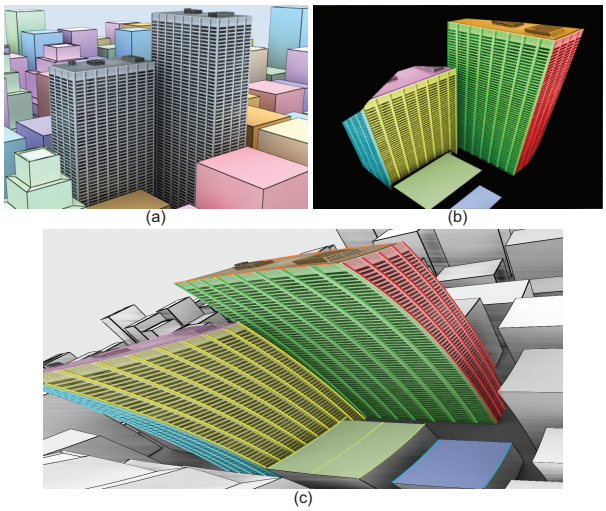

Figure 5. 3D reconstruction on the skyscraper scene. (a) A perspective image of the scene; (b) An XSlit image of the scene. The detected planes are highlighted in different colors. (c) The perspective rendering of our reconstruction.

a 3D plane and are properly clipped using the endpoints. The final recovered geometry is shown in Fig. 4(c). A total of 7 planes are detected in this example and we use different colors to highlight these planes. We also re-render the recovered faces using a perspective camera as shown in Fig. 4(d). Our solution is able to robustly and accurately recover most building faces from a single XSlit image.

Next, we generate a more complex and realistic skyscraper scene. We construct this scene using Autodesk 3ds Max (usa.autodesk.com/3ds-max). We place 2 building models from Evermotion (www.evermotion.org) at the center of the scene and surround them with multiple axis-aligned boxes of various heights. This emulates a complex MW scene. In this example, we use the image synthesis technique for producing an XSlit panorama [25, 30]: we translate a perspective camera horizontally from left to right with constant speed and then stitch linearly varying columns of pixels. Each view is rendered at resolution of 300×600 and the perspective camera views the scene from top to down and tilted by 15° around the z axis. The synthesized XSlit panorama resembles a POX-Slit image. To reduce aliasing, we use a small translation step so that the neighboring views have small parallax (ranging from 1 to 4 pixels depending on object depth).

Next, we apply Canny edge detection to locate curve segments and discard short ones below a pre-defined threshold. We then fit conics to the detected curves. Specifically, we detect a total of 35 conics which result in 8 CCPs and therefore 8 planes. We are unable to recover the building blocks completely or largely occluded by the two skyscrapers. Similar to the box scene, we use the approach in Sec. 3.3 to reconstruct the 8 planes from their CCPs. The 35 line segments are also mapped back for clipping the planes.

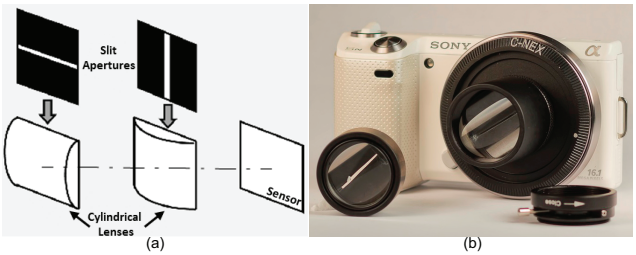

Figure 6. Our prototype XSlit Camera. (a) We use two layers of cylindrical lenses, each with a slit aperture; (b) We mount the XSlit lens on an interchangeable lens camera.

The final recovered planes are shown in different colors in Fig. 5(c). We re-render the reconstructed geometry using a perspective camera and compare it with the ground truth (Fig. 5(a) and (b)). Our results show that the XSlit-camera-based solution can robustly handle complex scenes.

4.2. Real Scene Experiments

Finally, we have constructed a prototype XSlit camera by modifying a commercial interchangeable lens camera (Sony NEX-5N). We replace its lens with a pair of cylindrical lenses, each using two slit apertures as shown in Fig. 6. We choose to modify an interchangeable lens camera rather than an SLR is that it has a shorter *flange focal distance* (FFD), *i.e.*, the distance between the sensor and the lens mount. For a 3D line to appear sufficiently curved, the line should span a large depth range w.r.t. the image plane (Eqn. (7)). This indicates that we need to put the camera closer to the objects as well as use lenses with a large field-of-view and a smaller focal length. The mirror-free interchangeable camera has a much shorter FFD than SLRs and therefore highly suitable. In our prototype, we use use two cylindrical lenses, one (closer to the sensor) with focal length 25mm and the other 75mm. The use of slim slit apertures, however, leads to low light efficiency: the XSlit image appears noisy. To calibrate the XSlit camera, we use a pattern of five lines and use an auxiliary perspective camera to determine line positions and orientations. We then conduct curve fitting for recovering the XSlit intrinsics.

In Fig. 7, we construct a scene composed of the parallel lines lying on two different planes. We also put a minifigure between the two planes. When viewed by a perspective camera, the lines appear nearly identical: although they intersect at a common vanishing point, it is difficult to tell if they belong to different planes, as shown in Fig. 7(b). In contrast, these lines are apparently different in our XSlit camera image, as shown in Fig. 7(c): they exhibit different curviness and one can directly tell that they do not belong to the same plane. Next, we apply the conic fitting and CCP detection methods on these curves and we are able to identify one XVP and two CCPs. Fig. 7(d) maps the recovered planes (highlighted in red and green) back onto

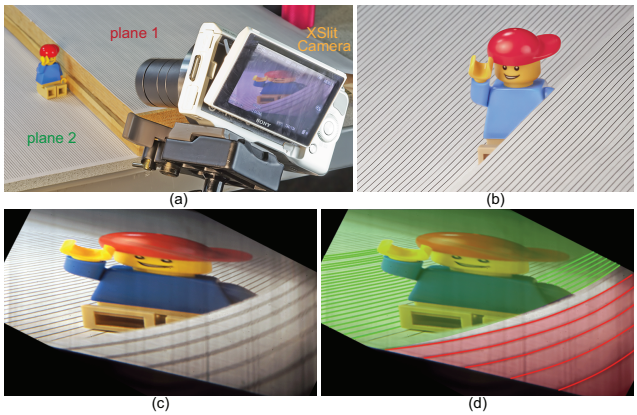

Figure 7. A simple scene consists of two sets of parallel lines on different planes. (a) Scene acquisition; (b) A perspective image; (c) An XSlit image; (d) We detect the two planes (highlighted in red and green) using the XSlit image.

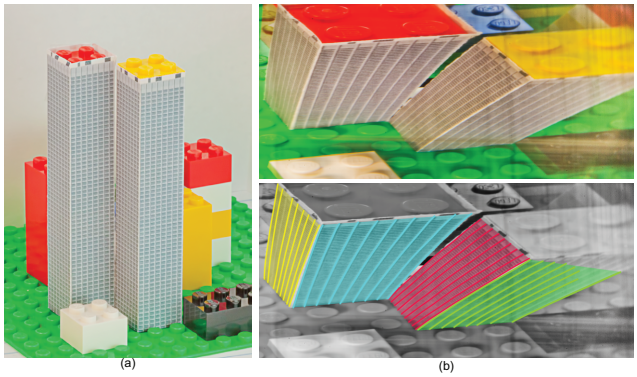

Figure 8. An MW Scene constructed from LEGO® bricks. (a) Image captured by a commodity camera; (b) Top: image captured by our XSlit camera; bottom: our recovered four building faces.

the acquired XSlit image. Our results align well with the ground truth.

Fig. 8 shows another real scene constructed using LEGO® building bricks, each of dimension $0.7" \times 0.7" \times 3.5"$. We capture the scene from a top-down perspective and orient the XSlilt camera to guarantee it observes enough curviness of vertical parallel lines. Without any processing, the XSlit image shows that the four groups of parallel lines exhibit different curviness. We detect 34 curves in the XSlit image and conduct our algorithm finding the XVPs and CCPs. Due to defocus blurs, some curves (yellow region) were missed or truncated (top of the green region). Nevertheless, our approach is still able to recover 4 faces (planes) of two buildings.

5. Conclusions and Discussions

We have presented a new solution for MW scene reconstruction from a unique perspective of non-pinhole imaging. Our solution directly resolves parallel line ambiguity by utilizing a unique class of image features in XSlit images,

i.e., the XSlit Vanishing Point (XVP) and Coplanar Common Point (CCP), for grouping coplanar parallel lines. Our main contribution is a new theory that shows each group of coplanar parallel lines will intersect at an XVP and a CCP in their XSlit image and its geometry can be directly recovered from the XVP and CCP. We have further developed a robust algorithm for automatically distinguishing XVPs from CCPs in complex MW scenes and validated our solution on both synthetic and real XSlit images.

Our solution relies on accurately detecting curves and fitting conics to locate XVPs and CCPs. If a captured curve is too short or too straight, our conic fitting scheme can introduce large errors and therefore generate incorrect XVPs and CCPs. One possible solution is to fit all conics together instead of each individual one by imposing the XSlit intrinsic constraint. As our main contribution is in theory, nearly all our experiments were conducted on synthetic scenes where we can easily modify the XSlit geometry to handle different scene configures. Our real XSlit camera, however, has limited capability on controlling slit distance and orientations. For example, due to the small distance between the slits, only 3D lines lying close to the camera will appear sufficiently curved. Further, we have to use relatively large slit apertures for maintaining light efficiency. As a result, they lead to a shallow depth-of-field and the conics appear blurry. In the future, we plan to explore dynamic aperture controls using LCoS [19] or other physical implementations of XSlit camera, *e.g.*, the rolling shutter camera [18].

Finally, our prototype XSlit camera may also benefit several other vision tasks. For example, by translating or rotating the camera, we can acquire the scene using a sequence of XSlit cameras. Using two XSlits, we can directly apply stereo matching as their images satisfy the Seitz condition [8, 25, 20]. Another potential application of our design is to construct non-pinhole light sources. In particular, we plan to explore the dual of the XSlit camera, *i.e.*, the XSlit light source. For example, we can combine an area light source with two cylindrical lenses for creating a prototype XSlit light source. Our theory in this paper can be used to guide new XSlit-based shape-from-shading and shape-from-shadow algorithms.

Acknowledgements

We thank Mohit Gupta and Shree Nayar for their invaluable suggestions and guidance. This project was partially supported by the National Science Foundation under grants IIS-CAREER-0845268 and IIS-RI-1016395, and by the Air Force Offce of Science Research under the YIP Award.

References

[1] O. Barinova, V. Konushin, A. Yakubenko, K. Lee, H. Lim, and A. Konushin. Fast automatic single-view 3D reconstruction of urban scenes. In *ECCV*, 2008.

[2] V. Caglioti and S. Gasparini. On the localization of straight lines in 3D space from single 2D images. In *CVPR*, 2005.

[3] J. M. Coughlan and A. L. Yuille. Manhattan world: Compass direction from a single image by bayesian inference. In *ICCV*, 1999.

[4] A. Criminisi, I. Reid, and A. Zisserman. Single view metrology. *IJCV*, 40(2):123–148, Nov. 2000.

[5] E. Delage, H. Lee, and A. Y. Ng. Automatic single-image 3D reconstructions of indoor manhattan world scenes. In *ISRR*, 2005.

[6] E. Delage, H. Lee, and A. Y. Ng. A dynamic bayesian network model for autonomous 3D reconstruction from a single indoor image. In *CVPR*, 2006.

[7] Y. Ding, J. Yu, and P. Sturm. Recovering specular surfaces using curved line images. In *CVPR*, 2009.

[8] D. Feldman, T. Pajdla, and D. Weinshall. On the epipolar geometry of the crossed-slits projection. In *ICCV*, 2003.

[9] A. Flint, C. Mei, D. Murray, and I. Reid. A dynamic programming approach to reconstructing building interiors. In *ECCV*, 2010.

[10] Y. Furukawa, B. Curless, S. M. Seitz, and R. Szeliski. Manhattan-world stereo. In *CVPR*, 2009.

[11] Y. Furukawa, B. Curless, S. M. Seitz, and R. Szeliski. Reconstructing building interiors from images. In *ICCV*, 2009.

[12] R. Hartley and A. Zisserman. *Multiple View Geometry in Computer Vision*. Cambridge University Press, NY, USA, second edition, 2003.

[13] D. Hoiem, A. Efros, and M. Hebert. Geometric context from a single image. In *ICCV*, 2005.

[14] D. Hoiem, A. A. Efros, and M. Hebert. Automatic photo pop-up. In *ACM SIGGRAPH*, 2005.

[15] J. Kosecka and W. Zhang. Video compass. In *ECCV*, 2002.

[16] D. C. Lee, M. Hebert, and T. Kanade. Geometric reasoning for single image structure recovery. In *CVPR*, 2009.

[17] M. Levoy and P. Hanrahan. Light field rendering. In *ACM SIGGRAPH*, pages 31–42, 1996.

[18] M. Meingast, C. Geyer, and S. Sastry. Geometric models of rolling-shutter cameras. *CoRR*, 2005. http://arxiv.org/abs/cs/0503076.

[19] H. Nagahara, C. Zhou, T. Watanabe, H. Ishiguro, and S. K. Nayar. Programmable aperture camera using LCoS. In *ECCV*, 2010.

[20] T. Pajdla. Epipolar geometry of some non-classical cameras. In *Proc. of Computer Vision Winter Workshop*, Slovenian Pattern Recognition Society, pages 223–233, 2001.

[21] J. Ponce. What is a camera? In *CVPR*, 2009.

[22] A. Saxena, S. H. Chung, and A. Y. Ng. Learning depth from single monocular images. In *NIPS*. 2005.

[23] A. Saxena, M. Sun, and A. Y. Ng. Make3D: Learning 3D scene structure from a single still image. *IEEE TPAMI*, 31(5):824 –840, May 2009.

[24] G. Schindler and F. Dellaert. Atlanta world: an expectation maximization framework for simultaneous low-level edge grouping and camera calibration in complex man-made environments. In *CVPR*, 2004.

[25] S. M. Seitz and J. Kim. The space of all stereo images. *IJCV*, 48(1):21–38, June 2002.

[26] R. Swaminathan, A. Wu, and H. Dong. Depth from distortions. In *OMNIVIS*, 2008.

[27] J. Yu and L. McMillan. General linear cameras. In *ECCV*, 2004.

[28] J. Yu, L. McMillan, and P. Sturm. Multi-perspective modelling, rendering and imaging. *Computer Graphics Forum*, 29(1):227–246, 2010.

[29] W. Zhang and J. Kosecka. Extraction, matching and pose recovery based on dominant rectangular structures. In *IEEE International Workshop on Higher-Level Knowledge in 3D Modeling and Motion Analysis*, Oct. 2003.

[30] A. Zomet, D. Feldman, S. Peleg, and D. Weinshall. Mosaicing new views: the crossed-slits projection. *IEEE TPAMI*, 25(6):741–754, June 2003.