Chapter #

Dynamic Pushbroom Stereo Vision

Dynamic Pushbroom Stereo Vision for Surveillance and Inspection

Zhigang Zhu¹, George Wolberg¹, Jeffery R. Layne²

¹Department of Computer Science, The City College of New York, New York, NY 10031
²Air Force Research Laboratory, 2241 Avionics Circle, WPAFB, Ohio 45433-7318

Abstract: We present a dynamic pushbroom stereo geometry model for both 3D reconstruction and moving target extraction in applications such as aerial surveillance and cargo inspection. In a dynamic pushbroom camera model, a “line scan camera” scans across the scene. Both the scanning sensor and the objects in the scene are moving, and thus the image generated is a “moving picture” with one axis being space and the other being time. We study the geometry under a linear motion model for both the sensor and the object, and we investigate the advantages of using two such scanning systems to construct a dynamic pushbroom stereo vision system for 3D reconstruction and moving target extraction. Two real examples are given using the proposed models. In the first application, a fast and practical calibration procedure and an interactive 3D estimation method are provided for 3D cargo inspection with dual gamma-ray (or X-ray) scanning systems. In the second application, dynamic pushbroom stereo mosaics are generated by using a single camera mounted on an airplane, and a unified segmentation-based stereo matching algorithm is proposed to extract both 3D structures and moving targets from urban scenes. Experimental results are given.

Key words: video mosaicing, motion analysis, stereo vision, 3D reconstruction, moving target extraction

1. INTRODUCTION

Pushbroom images (or mosaics, when generated from video sequences) with parallel-perspective projections are very suitable representations for surveillance and/or security applications where the motion of the camera has a dominant translational direction. Examples include satellite pushbroom
imaging (Gupta & Hartley, 1997), airborne video surveillance (Zhu, et al, 2001, 2004), 3D reconstruction for image-based rendering (Chai & Shum, 2000), road scene representations (Zheng & Tsuji, 1992; Zhu & Hanson, 2004), under-vehicle inspection (Dickson, et al, 2002; Koschan, et al, 2004), and 3D measurements of industrial parts by an X-ray scanning system (Gupta, et al, 1994; Noble, et al, 1995). A pushbroom image/mosaic is a parallel-perspective image which has parallel projection in the direction of the camera’s motion and perspective projection in the direction perpendicular to that motion. A pair of pushbroom stereo images/mosaics can be used for both 3D viewing and 3D reconstruction when they are obtained from two different oblique viewing angles. An advantageous feature of the pushbroom stereo model is that depth resolution is independent of depth (Chai & Shum, 2000; Zhu, et al 2001). Therefore, better depth resolution can be achieved than with perspective stereo or the recently developed multi-perspective stereo with circular projection (Peleg, et al, 2001; Shum & Szeliski, 1999; Klette, et al, 2001), given the same image resolution. We note that multi-perspective stereo with circular projection that is based on wide-baseline line cameras can achieve very accurate depth resolution for far-range airborne scenes (Klette, et al, 2001). However, in such a configuration depth resolution is still proportional to the square of depth. Therefore, the depth accuracy varies greatly for cases when ranges of depths are large, in such applications as cargo inspection or ground robot surveillance. In addition, the circular motion that is required is not the best form for scanning long cargo containers, or walking through large scale 3D scenes.

Using pushbroom stereo images/mosaics for 3D viewing and/or 3D reconstruction has been studied for satellite imaging, airborne video mosaicing, under-vehicle inspection, street scene modeling, and industrial quality assurance. However, as far as we know, previous work on the aforementioned stereo panoramas (mosaics) only deals with static scenes. Most of the approaches for moving target tracking and extraction, on the other hand, are based on interframe motion analysis and expensive layer extraction (Zhou and Tao 2003; Xiao and Shah 2004; Collins, 2003). In security and inspection applications, quite a few X-ray or gamma-ray cargo inspection systems have been put to practical uses (Hardin, 2002; Hardin, 2004; Hitachi, 2004). In the past, however, cargo inspection systems have only had two-dimensional capabilities, and human operators made most of the measurements. No moving target detection capability has been explored. If we could build an accurate geometry model for a X-ray/gamma-ray imaging system, which turns out to be a linear pushbroom scanning sensor, accurate three-dimensional (3D) and possible dynamic measurements of objects inside a cargo container can be obtained when two such scanning
systems with different scanning angles are used to construct a dynamic linear pushbroom stereo system. The 3D and motion measurements of targets add more value to today’s cargo inspection techniques, as indicated in some online reports (Hardin, 2002; Hardin, 2004; Hitachi, 2004).

In this work, we present a unified geometric model for a dynamic linear pushbroom stereo system for applications where the movements of both sensors and objects are involved. This raises very interesting problems since a pushbroom image is a spatio-temporal image with moving viewpoints, and thus the behavior of a moving object in the pushbroom image will be more complicated than in a conventional 2D spatial image captured in a single snapshot. We study the accurate geometric model of a linear pushbroom sensor when both the sensor and a 3D point move in 3D space at constant velocities. Then, we discuss why this model can be used for real applications where the motion of the objects can be well-approximated as piecewise linear within short periods of time.

In particular, we will study two examples using this model. In the first application, issues on 3D measurements using a linear pushbroom stereo system are studied for X-ray/gamma-ray cargo inspection. The closest work to ours is the x-ray metrology for industrial quality assurance (Noble, et al, 1995). However, to our knowledge, our research presents the first piece of work in using linear pushbroom stereo for 3D gamma-ray or X-ray inspection of large cargo containers. Furthermore, the proposed dynamic pushbroom stereo model enables moving target extraction within cargo containers. In our study, we use the gamma-ray scanning images provided by the Science Applications International Corporation (SAIC) (Orphan, et al, 2002). However, this does not imply an endorsement of this gamma-ray technology over others, for example, the X-ray technologies. In fact, the algorithms developed in this work can be used for pushbroom images acquired by X-ray or other scanning approaches as well.

In the second application, we study a dynamic pushbroom stereo mosaic approach for representing and extracting 3D structures and independent moving targets from urban 3D scenes. Our goal is to acquire panoramic mosaic maps with motion tracking information for 3D (moving) targets using a light aerial vehicle equipped with a video camera flying over an unknown area for urban surveillance. In dynamic pushbroom stereo mosaics, independent moving targets can be easily identified in the matching process of stereo mosaics by detecting the “out-of-place” regions that violate epipolar constraints and/or give 3D anomalies. We propose a segmentation-based stereo matching approach with natural matching primitives to estimate the 3D structure of the scene, particularly the ground structures (e.g., roads) on which humans or vehicles move, and then to identify moving targets and to measure their 3D structures and movements.
The rest of the text is organized as follows. In Section 2, the principle of the dynamic linear pushbroom imaging and then the model of dynamic pushbroom stereo is described. Section 3 discusses sensor calibration and 3D measurement issues for the gamma-ray cargo inspection application. Section 4 presents the stereo mosaicing approach to generate pushbroom stereo images from a single video sequence of a urban scene, and then to extract both 3D and moving targets from a pushbroom stereo pair. Section 5 provides some concluding remarks and discussion.

2. DYNAMIC PUSHBROOM STEREO GEOMETRY

We start with the proposed dynamic linear pushbroom imaging geometry for both static and dynamic scenes. Then, we study pushbroom stereo geometry for static scenes followed by dynamic pushbroom stereo geometry for dynamic scenes.

2.1 Dynamic linear pushbroom geometry

A 3D point at time \( t = 0 \) is denoted as \( P = (x, y, z) \) in a world coordinate system \( o-xyz \) (Figure 1). It is viewed by a moving camera \( O_c-X_cY_cZ_c \) with an optical axis \( Z_c \) and a focal length \( f \), but only points on the \( O_cY_cZ_c \) plane are recorded on a 1D scan line in the direction of \( v \). The center of the linear image in the \( v \) direction is defined by a vertical offset \( p_v \). The relation between the world coordinate system and the camera coordinate system located at the time \( t = 0 \) is represented by a rotational matrix \( R \) and a translational vector \( T \). Both the camera and the point translate at constant velocities, \( V_c \) and \( V_o \) respectively. Both of them are represented in the world coordinate system. The relative motion between them is denoted as \( V \) in the camera coordinate system. Therefore, we have

\[
V = (V_x, V_y, V_z)^T = R(V_c + V_o)
\] (1)

Using the linear pushbroom model proposed in Gupta and Hartley (1997), we have our dynamic linear pushbroom camera model:

\[
\begin{bmatrix}
    u \\
    wv \\
    w
\end{bmatrix} = \begin{bmatrix}
    1 & 0 & 0 & 1/V_x & 0 & 0 \\
    0 & f & p_v & -V_y/V_x & 1 & 0 \\
    0 & 0 & 1 & -V_z/V_x & 0 & 1
\end{bmatrix} \begin{bmatrix}
    x \\
    y \\
    z \\
    1
\end{bmatrix}
\] (2)
Note that the form is the same, but here we assume that both the camera and the point are moving during imaging, while in Gupta and Hartley (1997) only the camera moves at a constant velocity.

![Dynamic linear pushbroom imaging geometry](image.png)

Figure #1. Dynamic linear pushbroom imaging geometry. A 3D point at time \( t = 0 \) is denoted as \( P = (x,y,z)^T \) in the world coordinate system \( o-xyz \). The scene is viewed by a moving camera \( O_c-X_cY_cZ_c \) with an optical axis \( Z_c \), but only those points that pass the \( O_cY_cZ_c \) plane are recorded on a 1D scan line in the direction of \( v \). Both the camera and the point translate at constant velocities, \( V_c \) and \( V_o \) respectively.

The linear pushbroom model for static scenes with general parameters has been studied thoroughly in Gupta and Hartley (1997), including camera calibration, the epipolar geometry and fundamental matrix for a pushbroom stereo system. In our work, we are more interested in investigating the behavior of a moving target under a linear pushbroom imaging system, therefore extend the model to a dynamic version. Since it will be very complicated with a general model, for simplify the discussion we make the following assumptions of the orientation and motion direction of the camera, i.e.,

\[
R = \begin{pmatrix}
\cos \theta & 0 & -\sin \theta \\
0 & 1 & 0 \\
\sin \theta & 0 & \cos \theta
\end{pmatrix}, \quad V_c = (S,0,0)
\] (3a)

That is, the camera moves in the direction of the \( x \)-axis of the world coordinate system, and there are no tilt and roll angles between the two coordinate systems; the only rotation is around the \( y \)-axis. Suppose that \( T = (T_x,T_y,T_z)^T \) and
then Eq. (2) can be written as our dynamic linear pushbroom imaging model:

\[
u = \frac{(x-T_x)-(z-T_z)\tan \theta}{S+W_x-W_z\tan \theta}, \quad v = f'\cos \theta \frac{y-T_y-W_yu}{z-T_z-W_zu} + p_v\quad (4)
\]

Note that the linear pushbroom camera system has parallel projection in the \( u \) direction, but has perspective projection in the \( v \) direction. Recall that \((x,y,z)\) is the “initial” point location in the word at the time \( t = 0 \). The current location seen in the image (at the time \( t = u \)) can be calculated as

\[
P_t = (x, y, z)' - u(W_x, W_y, W_z)'
\]

However, the moving point does not have to keep the linear motion starting from time \( t=0 \) to make Eq. (4) valid. The time \( t = u \) is the only time it is seen in the current temporal-spatio \( u-v \) image, so the initial location is only an assumed location that will be useful in the dynamic pushbroom stereo model in the Section 2.3.

### 2.2 Linear pushbroom stereo for static scenes

First, we want to look at a point in a static scene, i.e., when \( V_0 = (W_x, W_y, W_z)' = 0 \). If we have two linear pushbroom scanning systems that have two different sets of parameters \((\theta_k, T_k, f_k, p_{vk})\), \( k = 1, 2 \), we have from Eq. (4) the relations for a pair of correspondence points \((u_k, v_k)\), \( k=1,2 \):

\[
u_k = \frac{x-T_{xk}-(z-T_{zk})\tan \theta_k}{S_k}, \quad (k=1,2)
\]

\[
v_k = f_k\cos \theta_k \frac{y-T_{yk}}{z-T_{zk}} + p_{vk}
\]

Using sensor calibration (Section 3.1), we can find these parameters. Therefore, the depth of the point can be recovered as

\[
z = \frac{d-d_0}{\tan \theta_1 - \tan \theta_2}
\]
where

\[ d = S_2 u_2 - S_1 u_1 \]  

(8)

is defined as the visual displacement (measured in metric distance) of the point \((x,y,z)\) as observed in the pair of stereo images, and

\[ d_0 = (T_{x1} - T_{x1} \tan \theta) - (T_{x2} - T_{x2} \tan \theta_2) \]

is the fixed offset between two images. Note that Eq. (7) is acquired by only using the \(u\) coordinates of the stereo images, and the depth of any point is proportional to its visual displacement in the stereo pair. Thus, the depth resolution in a linear pushbroom stereo system is independent of depth. The epipolar geometry can be derived from Eq. (6), which has been proved to be of hyperbolic curves (Gupta and Hartley, 1997).

2.3 Linear pushbroom stereo for dynamic scenes

For a dynamic scene, with a moving point \(V_o = (W_x, W_y, W_z)^T \neq 0\), we need to estimate not only the depth \(z\) but also the motion \(V_o\) of the dynamic point. Without loss of generality, we assume \(T = (T_x, T_y, T_z)^T = 0\), \(p_x = 0\), i.e., the camera is at the origin of the world coordinate system at time \(t = 0\), and we have to find the center of perspective image by calibration. Therefore we have

(Figure 2. Dynamic linear pushbroom stereo geometry. The \(y\) axis points out of the paper.)
Assume that two linear pushbroom scanning systems with two different sets of parameters \((\theta_k, f_k)\) scan the scene at the same time, starting at the same camera location (Figure 2). For a pair of correspondence points \((u_k, v_k)\), \(k=1,2\), of a 3D moving point \(P (x,y,z)\), we have the following dynamic linear pushbroom stereo model:

\[
 u_k = \frac{x - z \tan \theta_k}{S + W_x W_z \tan \theta_k},
 v_k = f \cos \theta_k \frac{y - W_y u_k}{z - W_z u_k}, \quad k=1,2
\]  

(10)

Hence the depth can be represented as

\[
 z = \frac{S + W_x}{\tan \theta_1 - \tan \theta_2} d_u + \frac{u_1 \tan \theta_1 - u_2 \tan \theta_2}{\tan \theta_1 - \tan \theta_2} W_z
\]  

(11)

where

\[
d_u = u_2 - u_1
\]  

(12)

is the temporal displacement (measured in numbers of scan lines) as observed in the stereo pair. In the dynamic linear pushbroom stereo model, we only need to assume that the point \(P\) moves at a constant velocity \(V_0\) from time \(t = u_1\) to time \(t = u_2\). These are the only two time instants that points can be seen by the stereo system (Figure 2). The locations of the points at these two time instants are

\[
P_k = (x, y, z)^t - u_k (W_x, W_y, W_z)^t, \quad k = 1, 2
\]  

(13)

In real applications with a moving video camera, if we extract two scanlines in each 2D perspective image to simulate the two scanners with two angles (Zhu, et al 2004), we will have \(F = f \cos \theta_k\), \(k=1,2\), where \(F\) is the real focal length of the perspective camera, and its optical axis is parallel to the \(z\) axis of the world (Figure 2).

There are two remaining questions: how can we know a point is a dynamic point, and then how can we estimate the motion velocity of the point? We know that if the point is static, the correspondence point of \((u_1, v_1)\) in the second image, \((u_2, v_2)\), will on its parabolic epipolar line. Or in the
ideal model of Eq. (10) with $F = f \cos \theta_k$, $k=1,2$, on a horizontal straight line (i.e., $v_1 = v_2$). Therefore, in the general case (i.e. $W_y \neq 0$ or $W_z \neq 0$), a moving point will violate the epipolar geometry of the static scene condition.

Note that if the point moves in $xoy$ plane, the correspondence point will be on a straight line $v_2 - v_1 = -FW_y(u_2 - u_1)/z$ instead of on the parabolic epipolar curve. However, when the motion of the point is in the direction of the camera’s motion, i.e., both $W_y$ and $W_z$ are zeros, the moving point will also obey the epipolar geometry of static assumption. But in this special case we can use a depth anomaly test (for aerial surveillance where moving objects are on the ground) or depth difference constraints (for known objects), as below.

In summary, the velocity of the dynamic point can be estimated as the following two cases, depending on if $W_z=0$ or not.

**Case 1.** The point moves in the $xoy$ plane (i.e., $W_z=0$). The following two steps will be used in estimating the motion of the point:

1. If $z$ can be obtained from its surroundings, or if the depth difference between two points (e.g., “height” of an object) is known, then $W_x$ can be calculated from Eq. (11);
2. Given $z$, both $y$ and $W_y$ can be obtained from the equations for $v_k$ ($k=1,2$) in Eq. (10) with a pair of the linear equations for $v_1$ and $v_2$.

**Case 2.** General case ($W_z \neq 0$). We also have two steps to fulfil the task:

1. If three points on the same moving object have the same depth (or two points have known depths), then $z$, $W_x$ and $W_y$ can be obtained from Eq. (11) with a linear equation system, and then $x$ from an equation for $u_k$ ($k = 1$ or 2) in Eq. (10);
2. Given $W_x$, $x$ and $z$, both $y$ and $W_y$ can be obtained from the equations for $v_k$ ($k=1,2$) in Eq. (10) with a pair of linear equations for $v_1$ and $v_2$.

## 3. GAMMA-RAY LINEAR PUSHBROOM STEREO

The system diagram of the gamma-ray cargo inspection system is shown in Figure 3b. A 1D detector array of 256 NaI-PMT probes counts the gamma-ray photons passing through the vehicle/cargo under inspection from a gamma-ray point source. Either the vehicle/cargo or the gamma-ray system (the source and the detector) moves in a straight line in order to obtain a 2D scan of gamma-ray images. The geometry of the system is shown in Figure 3a, which is essentially the same as in Figure 1.
A dual-scanning system is a linear pushbroom stereovision system. It can be constructed with two approaches: two linear pushbroom scanning sensors with different scanning angles, or a single scanning sensor to scan the same cargo twice with two different scanning directions. The first approach can be
used to detect moving targets inside a cargo container. Figure 4 shows two real gamma-ray images, with different scanning angles – ten and twenty degrees, respectively. Each image has a size of 621x256 pixels, i.e., 621 scans of the 256-pixel linear images. For static scene points, depths (z) can be calculated by using Eq. (7), and further x and y coordinates can be calculated by Eq. (6). Note that here we assume that the two scans are independent of each other and thus have two different sets of imaging parameters. Two important steps are described in the following for 3D measurements in cargo inspection: sensor calibration and stereo matching.

### 3.1 Sensor Calibration

For each scanning setting, the following parameters are required for 3D estimation: the focal length \( f \), the image center \( p_i \), the scanning angle \( \theta \), the scanning speed \( S \), and the initial sensor location \( (T_x, T_y, T_z) \). In order to fulfill this task, we need to know a set of 3D points \( \{(x_i, y_i, z_i)\} \) and their corresponding image points \( \{(u_i, v_i)\}, \) for \( i = 1, 2, \ldots, N \). Our calibration method only needs to know the dimension of the container, which is

\[
\text{length}(x) \times \text{height}(y) \times \text{depth}(z) = 20 \times 8 \times 8 \text{ (ft}^3\text{).}
\]

Then we locate the 8 vertices of the rectangular container (refer to Figure 3a) in each gamma-ray image by manually picking up the 8 corresponding image points.

An interesting property of the linear pushbroom sensor is that the two equations in Eq. (4) can work independently (with \( W_x = W_y = W_z = 0 \)). Therefore, in calibrating the sensor, we first obtain the “parallel projection parameters” from \( u \) and then the “perspective projection parameters” from \( v \). The parallel projection equation can be turned into a linear equation with three unknowns, i.e., \( S, \tan \theta \) and \( T_x - T_z \tan \theta \):

\[
u_i S + z_i \tan \theta + (T_x - T_z \tan \theta) = x_i
\]  

Given more than three pairs of points \( (i=1, 2, \ldots, N \text{ where } N \geq 3) \), we can solve the linear system to find the three unknowns by using the least square method. Similarly, the perspective equation leads to a linear equation with five unknowns, i.e. \( f, fT_y, p_v, p_zT_z \) and \( T_z \):

\[
(y_i \cos \theta)f - \cos \theta(fT_y) + z_i p_v - (p_zT_z) + v_i T_z = v_i z_i
\]  

With the known \( \theta \) and given more than five pairs of points \( (i=1, 2, \ldots, N \text{ where } N \geq 5) \), we can solve the linear equation system. Note that from Eq.
(14) we can only find the values of the speed $S$ and the angle $\theta$ and a combined parameter $T_x-T_z \tan \theta$. Nevertheless, this is sufficient for obtaining the depths of points using Eq. (7). Table 1 shows the results of the “parallel parameters” for the two settings corresponding to the two images in Figure 4. All the rest of the parameters, including $T_x$, can be obtained after solving Eq. (15), in order to calculate the $x$ and $y$ coordinates of 3D points by using Eq. (6). Table 2 shows the “perspective parameters” and the $T_x$ values for the two settings.

Table #1. Parallel projection parameters

<table>
<thead>
<tr>
<th>Images</th>
<th>$S$ (ft/pixel)</th>
<th>$\tan \theta$</th>
<th>$\theta$ (degrees)</th>
<th>$T_x-T_z \tan \theta$</th>
</tr>
</thead>
<tbody>
<tr>
<td>10-degrees</td>
<td>0.04566</td>
<td>0.16552</td>
<td>9.3986</td>
<td>-7.283</td>
</tr>
<tr>
<td>20-degrees</td>
<td>0.04561</td>
<td>0.34493</td>
<td>19.031</td>
<td>-7.309</td>
</tr>
</tbody>
</table>

Table #2. Perspective projection parameters

<table>
<thead>
<tr>
<th>Images</th>
<th>$F$ (pixels)</th>
<th>$T_y (ft)$</th>
<th>$p_x (pixels)$</th>
<th>$p_z T_z$</th>
<th>$T_z (ft)$</th>
<th>$T_x (ft)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>10-degrees</td>
<td>441.24</td>
<td>-0.42881</td>
<td>17.787</td>
<td>-191.78</td>
<td>-15.141</td>
<td>-9.789</td>
</tr>
<tr>
<td>20-degrees</td>
<td>456.18</td>
<td>-0.41037</td>
<td>19.250</td>
<td>-198.03</td>
<td>-15.000</td>
<td>-12.48</td>
</tr>
</tbody>
</table>

Table #3. 3D measurements of the test points (all measurements are in feet)

<table>
<thead>
<tr>
<th>No</th>
<th>$x$</th>
<th>$y$</th>
<th>$z$</th>
<th>$dx$</th>
<th>$dy$</th>
<th>$dz$</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>-0.033</td>
<td>-0.179</td>
<td>-0.063</td>
<td>-0.033</td>
<td>-0.179</td>
<td>-0.063</td>
</tr>
<tr>
<td>1</td>
<td>20.033</td>
<td>-0.177</td>
<td>0.063</td>
<td>0.033</td>
<td>-0.177</td>
<td>0.063</td>
</tr>
<tr>
<td>2</td>
<td>19.967</td>
<td>-0.152</td>
<td>7.936</td>
<td>-0.033</td>
<td>-0.152</td>
<td>0.064</td>
</tr>
<tr>
<td>3</td>
<td>0.033</td>
<td>-0.204</td>
<td>8.064</td>
<td>0.033</td>
<td>-0.204</td>
<td>0.064</td>
</tr>
<tr>
<td>4</td>
<td>-0.033</td>
<td>7.787</td>
<td>-0.063</td>
<td>-0.033</td>
<td>-0.213</td>
<td>-0.063</td>
</tr>
<tr>
<td>5</td>
<td>20.033</td>
<td>7.856</td>
<td>0.063</td>
<td>0.033</td>
<td>-0.144</td>
<td>0.063</td>
</tr>
<tr>
<td>6</td>
<td>19.967</td>
<td>7.799</td>
<td>7.936</td>
<td>-0.033</td>
<td>-0.201</td>
<td>0.064</td>
</tr>
<tr>
<td>7</td>
<td>0.033</td>
<td>7.844</td>
<td>8.064</td>
<td>0.033</td>
<td>-0.156</td>
<td>0.064</td>
</tr>
</tbody>
</table>

Table 3 shows the 3D measurements using the image point pairs used for calibration between two views, the ten-degree and the twenty-degree images. The purpose is to show how accurate the pushbroom stereo model and the calibration results are. The numbers of the points listed in Table 3 are labeled in Figure 3a for comparison. For the container with a dimension of 20x8x8 ft$^3$, the average errors in depth $z$, length $x$ and height $y$ are 0.064 ft, 0.033 ft, and 0.178 ft, respectively, indicating that the pushbroom modeling and calibration is accurate enough for 3D measurements. Note that the accuracy of the estimation only reflects the errors in sensor modeling and calibration. No image localization errors have been included. The depth error $\delta z$ introduced by image localization error $\delta u$ can be estimated as the first derivative of $z$ with respect to $u$ using Eqs. (7) and (8), that is...
In Eq. (16) we assume that two scans share the same speed (i.e., $S_1 = S_2 = S$), which are almost true for our example in Figure 4 (see Table 1). In this example, one-pixel image localization error introduces an error of 0.254 ft in depth estimation, using the parameters in Table 1. Analysis in more details on the calibration accuracy can be found in Zhu, et al (2005a).

### 3.2 3D Measurements and Visualization

Fully automated 3D measurements of objects from gamma-ray radiographic images are difficult since the objects in the images are “transparent.” Some work has been reported in literature (e.g., Mayntz, et al, 2000) in using optical flow on X-ray fluoroscopy images for restoration of motion blur, but the motion parallax in their case is small. However, in our case of widely separated parallel viewing for 3D reconstruction two different views will give very different object adjacencies and occlusions. This is an important issue and will be our future work. In our current work, we have tested an interactive approach for stereo matching and visualization.

Our semi-automated stereo matching approach includes three steps: interactive point selection, automated matching, and interactive matching correction. Instead of generating a dense “depth” map from a pair of gamma-ray images, we have designed an interactive user interface for selecting and measuring objects of interest. For the automated stereo matching step, we use sum of square difference (SSD) criterion on normalized images.

Figure 4 shows the process of semi-automated stereo matching for the pair of the ten- and twenty-degree images. After a point in the first image is picked up by the user (marked by a red star in the first image of Figure 4), its match in the second image is automatically searched along the epipolar line of the pushbroom stereo, derived from Eq. (6). The search range is predetermined from Eq. (7) by using the knowledge that all the objects are within the cargo container. The size of the correlation window can be determined by the user interactively. We have tried different window sizes (3x3, 9x9, 11x11, etc.) and found that 11x11 was the best for this example. The automated matches are marked by blue stars in the second image of Figure 4.

After the program finds the automated matching points, the user could correct the match if necessary (marked by green stars in the second image of Figure 4). In Figure 4, most of the automated matches are “considered” to be correct where the green marks completely overlap the blue marks. The
points that are considered incorrect are those whose matches could be identified by human eyes but whose appearances are quite different between two images for automated matching. On the other hand, a few point matches that are considered to be “correct” might be incorrect; but, we have no way to correct them due to the large differences between two views (e.g., the point pair identified by arrows). In Figure 4, all eight vertices of the cargo container are selected for stereo matching as well as a few points around the boundary of a vehicle inside the cargo container. Note that the four of the eight points on the top of the container we select here are slightly different from the ones for calibration due to the requirements of an 11x11 window centered at each point.

Together with the stereo matching interface, the reconstructed 3D structures are rendered as wire frames in 3D. For each set of points that are selected for stereo matching, a connected 3D line-frame representation is generated (Figure 5). In Figure 5, the black rectangular frame is the reconstruction of the cargo container using the calibration image data for the ten-and twenty-degree images. The red line frame is generated from the 3D measurements by the automated stereo match algorithm. It is clearly shown that the automated stereo matches provide very good 3D measurements for the cargo container and the objects inside. With a 3D visualization, 3D measurements, for example, of sizes and shapes are made simple by using the most convenient views. Object measurements and identification will be our future work.

Figure #5. 3D measurements and visualization of objects inside the cargo container. The black rectangular frames show the cargo container constructed from the test data in Table 3. The red lines (with stars) show the 3D estimates from automated stereo matches, for the cargo container and an object inside.
4. DYNAMIC STEREO MOSAICS FROM VIDEO

In this section, we will see how we can use the dynamic linear pushbroom stereo model by generating pushbroom stereo mosaics from an aerial video sequence. Our goal is to acquire geo-referenced mosaic maps with motion tracking information for 3D (moving) targets using a light aerial vehicle flying over an unknown 3D urban scene with moving targets (vehicles). The motion of the camera has a dominant direction of motion (as on an airplane or ground vehicle), but 6 DOF motion is allowed.

4.1 Linear pushbroom mosaic generation

First, for introducing the principle, we assume the motion of a camera is an ideal 1D translation with constant speed, the optical axis is perpendicular to the motion, and the frames are dense enough. Then, we can generate two spatio-temporal images by extracting two columns of pixels (perpendicular to the motion) at the leading and trailing edges of each frame in motion (Figure 6). The mosaic images thus generated are parallel-perspective, which have perspective projection in the direction perpendicular to the motion and parallel projection in the motion direction. In addition, these mosaics are obtained from two different oblique viewing angles ($\theta_1$ and $\theta_2$) of a single camera’s field of view, so that a stereo pair of “left” and “right” mosaics captures the inherent 3D information. The geometry in this ideal case (i.e., 1D translation with constant speed) is the same as the linear pushbroom stereo model represented in Eq. (10), where $T = 0$, and the camera focal length $F = f \cos \theta_k$, $k=1,2$.

![Diagram of parallel-perspective pushbroom stereo mosaics](image)

*Figure 6. Principle of the parallel-perspective pushbroom stereo mosaics.*

In real applications, there are two challenging issues. The first problem is that the camera usually cannot be controlled with ideal 1D translation and
camera poses are unknown; therefore, camera orientation estimation (i.e., dynamic calibration) is needed. Bundle adjustment techniques (Triggs, et al, 2000) can be used for camera pose estimation, sometimes integrated with the geo-referenced data from GPS and INS when available. The second problem is to generate dense parallel mosaics with a sparse, uneven, video sequence, under a more general motion. The ray interpolation approach we proposed in Zhu, et al (2004) can be modified to generate a pair of linear pushbroom stereo mosaics under the obvious motion parallax of a translating camera.

Figure 7 shows how the modified ray interpolation approach works for 1D cases. The 1D camera has two axes – the optical axis ($Z$) and the $X$-axis. Given the known camera orientation at each camera location, one ray with a given oblique angle $\theta$ can be chosen from the image at each camera location to contribute to the parallel mosaic with this oblique angle $\theta$. The oblique angle is defined against the direction perpendicular to the mosaicing direction, which is the dominant direction of the camera path (Figure 7, same as in Figure 2). But the problem is that the “mosaiced” image with only those existing rays will be sparse and uneven (i.e., not linear) since the camera arrays are usually not regular and dense. Therefore, interpolated parallel rays between a pair of existing parallel rays (from two neighboring images) are generated by performing local matching between these two images so that we generate rays at dense and equal intervals along the mosaicing direction, i.e., with a linear pushbroom geometry as if from a moving camera with constant velocity. Note that the velocity is measured in meters per scanline (ray) instead of time. The assumption is that we can find at least two images to generate the parallel rays. Such interpolated rays are shown in Figure 7, where Ray $I_1$ is interpolated from Image A and Image B for a static point $P_1$ by back-projecting its corresponding pair in images A and B, and Ray $I_2$ is interpolated from Image C and Image D for a moving point $P_2$ in the same way. By assume that the moving point undergoes a linear motion between the two views (from $P_2^C$ to $P_2^D$), the interpolated ray captures its correct location ($P_2^I$) at that ray (i.e., “time”).

![Figure 7](#). Ray interpolation for linear pushbroom mosaicing with parallel projection.
The extension of this approach to 2D images is straightforward, and a region triangulation strategy similar to the one proposed in Zhu, et al (2004) can be applied here to deal with 2D cases. Since the process of parallel ray interpolation as a “temporal re-sampling” synthesizes each mosaic with a linear time axis in the direction of mosaicing, for both the static and moving objects, a pair of such mosaics satisfies dynamic linear pushbroom stereo geometry represented by Eq. (9). In the following text, we assume that pushbroom stereo mosaics have been generated and that the epipolar geometry obeys that of the linear pushbroom stereo under 1D translation. We will then focus on the method to perform both 3D reconstruction and moving target extraction from urban scenes from a pair of dynamic pushbroom stereo mosaics.

4.2 Segmentation-based stereo matching

Simple window-based correlation approaches do not work well for man-made scenes, particularly across depth boundaries and for textureless regions. An adaptive window approach (Kanade & Okutomi, 1991) has been proposed which selects at each pixel the window size that minimizes the uncertainty in disparity estimates in stereo matching. A nine window approach has also been proposed by Fusiello, et al (1997) in which the point in the right image with the smallest SSD error amongst the 9 windows and various search locations is chosen as the best estimate for the given point in the left image. Recently, color segmentation has been used as a global constraint for refining an initial depth map to get sharp depth boundaries and to obtain depth values for textureless areas (Tao, et al 2001), and for accurate layer extraction (Ke & Kanade 2002). In this text, we provide a segmentation-based approach using natural matching primitives to extract 3D and motion of the targets. The segmentation-based stereo matching algorithm is proposed particularly for the dynamic pushbroom stereo geometry to facilitate both 3D reconstruction and moving target extraction from 3D urban scenes. However, the proposed natural matching primitives are also applicable to more general scenes and other types of stereo geometry.

Dynamic pushbroom stereo mosaics provide several advantages for 3D reconstruction and moving target extraction (please refer to Section 2.3). The stereo mosaics can be aligned on a dominant plane (e.g., the ground), as in Zhu, et al, 2004. All the static objects obey the epipolar geometry, i.e., along the epipolar lines of pushbroom stereo. An independent moving object, on the other hand, either violates the epipolar geometry if the motion is not in the direction of sensor motion or at least exhibit 3D anomaly - hanging above the road or hiding below the road even if motion happens to be in the
same direction of the sensor motion (Zhu, et al, 2005b). With all these geometric constraints in mind, we propose a segmentation-based approach to integrate the estimation of 3D structure of an urban scene and the extraction of independent moving objects from a pair of dynamic pushbroom stereo mosaics. The approach starts with one of the mosaics, for example, the left mosaic, by segmenting it into homogeneous color regions that are treated as planar patches. We apply the mean-shift-based approach proposed by Comaniciu & Meer (2002) for color segmentation. Then, the stereo matching is performed based on these patches between two original color mosaics. The basic idea is to only match those pixels that belong to each region (patch) between two images in order to both produce sharp depth boundaries for man-made targets and to facilitate the searching and discrimination of the moving targets (each covered by one or more homogeneous color patches).

The proposed algorithm has the following five steps.

Step 1. Matching primitive selection. After segmenting the left image using the mean-shift method, homogenous color patches and then the natural matching primitives are extracted.

Step 2. Epipolar test. Using pushbroom epipolar geometry in stereo matching, static objects will find correct matches but moving objects will be outliers without correct “matches”.

Step 3. 3D anomaly test. After ground surface fitting (and road detection), moving objects in the same motion direction will exhibit wrong 3D characteristics (hanging above roads or hiding below roads).

Step 4. Motion extraction. Search matches for outliers (which could be moving objects) with a 2D and larger search range, or along the road directions (if available).

Step 5. 3D estimation. Using the dynamic pushbroom stereo method proposed in Section 2.3, the 3D structures and motion of moving objects could be derived.

In the following two subsections, we detail two important issues in the segmentation-based stereo matching approach: natural matching primitive selection, and an integrated analysis of 3D structure and motion for both static and moving targets.

4.3 Natural matching primitives

We use color segmentation to obtain natural matching primitives for both 3D reconstruction and moving target extraction. The selection and matching of the natural matching primitives includes the following five sub-steps.

(1) Segmentation and Interest point extraction. The left mosaic is segmented into homogeneous color regions using the mean-shift approach (Comaniciu & Meer 2002). We assume that each homogeneous color region
(patch) is planar in 3D. However, each planar surface in 3D may be divided into several color patches. Then the boundary of each region is traced as a close curve. All the neighborhood regions are also connected with the region in processing for further use. Finally we use a line fitting approach to extract interest points along the region’s boundary. The boundary is first fitted with connected straight-line segments using an iterative curve splitting approach. The connecting points between line segments are defined as interest points.

(2) **Natural window definition.** Each interest point \( p(u,v) \) of a region \( R \) in consideration will be used as the center of an \( m \times m \) rectangular window in the left mosaic. Only those points that are within the window, inside the regions, or on the boundary will be used for matching (Figure 8) in order to keep sharp depth boundaries. The window is defined as a *natural matching window*, and the set of pixels involved in matching is called a *natural matching primitive*. To facilitate the computation of correlation for stereo matching, we define a region mask \( M \) of size \( m \times m \) centered at that interest point such that

\[
M(i, j) = \begin{cases} 
1, & \text{if } (u + i, v + j) \in R \\
0, & \text{otherwise} 
\end{cases}
\]  

(17)

We changed the size \( m \) of the natural window depending on the sizes of the regions. In our experiments, we use \( m = 23 \) for large regions (with diameter \( \geq 23 \)) and \( m = 15 \) for small regions. We also want to include a few more pixels (1-2) around the region boundary (but not belonging to the region) so that we have sufficient image features to match. Therefore, a dilation operation will be applied to the mask \( M \) to generate a region mask covering pixels across the depth boundary. Figure 8 illustrates four such windows for the four interest points on the top region of the box. Note the yellow-shaded portions within each rectangular window, indicating that the pixels for stereo matching cover the depth boundaries.

![Figure 8. Natural matching primitive](image_url)
(3) **Natural window-based correlation.** Let the left and right mosaics be denoted as $I_1$ and $I_2$, respectively. The weighted cross-correlation, based on the natural window centered at $p(u,v)$ in the left mosaic, is defined as

$$C(d_u,d_v) = \frac{\sum_{i,j} M(i,j)I_1(u+i,v+j)I_2(u+i+d_u,v+j+d_v)}{\sum_{i,j} M(i,j)}$$

Note that we still carry out correlation between two color images (mosaics), but only on those interest points on each region boundary and only with those pixels within the region and on the boundaries. This equation works for both static objects when the searching of correspondences is along epipolar lines of the pushbroom stereo and also for moving targets when the searching should be in 2D and with a larger search range. In the real implementation, we first perform matches with epipolar constraints of the pushbroom stereo, and those without good matches will be treated as “outliers” for further examination to see whether or not they are moving objects.

*Figure #9.* An example of region matching results. The matches are marked as “X”, with corresponding colors.

*Figure #10.* An example of surface fitting results.
Figure 9 shows a real example of natural-window-based stereo matching result for a static object (top of a building). The 19 selected interest points and their correspondences are marked on the boundaries in the left and right images, respectively. One mismatch and a small error in match are also indicated on images.

4.4 Surface fitting and motion estimation

Assuming that each homogeneous color region is planar in 3D, we fit a 3D plane for each region after we obtain the 3D coordinates of the interest points of the region using the pushbroom stereo geometry (assuming that it is static, i.e., $W = 0$ in Eqs. (10) and (11)). Seed points ($\geq 3$) are selected for plane fitting based on their correlation values. Then, the 3D coordinates of all the interest points are refined by constraining them on the fitted plane. Then, using the 3D plane information the region in the left mosaic can be warped to the right image to evaluate the matching and the fitting. Figure 10 shows the results of fitting and back-projection of the fitted region onto the right image. The 15 seed interest points (out of 19) used for planar fitting are indicated on the left image as squares. Both the mismatch and the small error in the initial match are fixed. Note that an iterative approach could be applied here to refine the matches after the initial surface fitting by using the evaluation of the warping from the left to the right mosaics and also by using the occlusion constraints from neighborhood regions, which have been obtained in the region tracing step in Section 4.3. For example, the selection of the seed points for surface fitting can be refined by removing those points that could be on occluding boundaries after we check the initial 3D surface relations and adding some other points that have reliable matches after image warping evaluation. Neighborhood regions can also be merged into a single plane if they share the same planar equation, with some error tolerant range.

After region refinement and merging, large and (near) horizontal ground regions can be easily identified. It is also possible to analyze the shape of the ground regions to estimate road directions in which vehicles move. For those smaller neighborhood regions that happen to move in the same direction as the camera, it will have large depth differences from the surrounding ground regions when treated as static objects. This 3D anomaly can be used to identify those regions as moving objects. By assuming that their depths are the same as the surroundings, their motion parameters can be estimated using the method described in Section 2.3. For those “outliers” that do not find matches in the first pass (along epipolar lines), searching for matches can be performed along possible road directions (if obtained from the surrounding ground regions), or simply performed in a much larger 2D searching range.
**Figure #11.** Dynamic pushbroom stereo mosaics and depth map (a) stereo mosaics: left view in the green/blue channels and right view in the red channel of a RGB image for stereo viewing; (b) depth map.

**Figure #12.** Content-based 3D mosaic representation: results for a portion of the stereo mosaics marked in Figure 11: (a) left color mosaic; (b) right color mosaic; (c) and (d) left color labels and region boundaries; (e) depth map of static regions; (f) moving targets (motion: blue to red). Note how close the color label image to the original color image is.
4.5 Experimental Results

We have performed preliminary experiments for stereo matching and moving object extraction on pushbroom stereo mosaics from real video sequences. Figure 11a shows a pair of stereo mosaics from a video sequence that was taken when the airplane was about 300 meters above the ground. The viewing angles of the two parallel viewing directions for this pair of pushbroom mosaics are about 0.0 degrees (“left view”) and 1.4 degrees (“right view”), respectively. Figure 11b shows the corresponding dense depth map for the entire mosaic (about 4K*1.5K pixels), with sharp depth boundaries and dense depth values for buildings with large depth ranges (0-50 meters) and textureless surfaces.

Figure 12 shows the results of a close-up window indicated in the stereo mosaics in Figure 11a. In Figures 12a and b, the dynamic pushbroom stereo pair has both stationary buildings and ground vehicles moving in different directions. Figures 12c and d show the segmentation result of the left image in Figure 12a, where the color label image is shown in c and the region boundaries are shown in d. Note that a planar 3D region may be segmented into several color patches. Figure 12e shows the depth map of the static regions. Note that many regions, particularly those on top of each building are correctly merged, and the depth boundaries are clearly sharp and accurate. Figure 12f shows the matched moving targets marked on the left mosaiced image, in blue and red, respectively. The moving vehicles exhibit much larger motion magnitudes and obvious different motion directions from the camera’ motion direction.

5. CONCLUSIONS AND DISCUSSIONS

We have built the imaging geometry model of dynamic scenes observed by a linear pushbroom imaging sensor, in which only one scanline is obtained at each time instant. Examples of this kind of imaging can be find in applications such as X-ray/gamma-ray cargo inspection and airborne or ground surveillance where a moving sensor is applied, and the observed scenes include moving targets of interest. We generalize the linear pushbroom camera model proposed by Gupta & Hartley for satellite imaging to the proposed dynamic linear pushbroom imaging model to deal with close-range scenes and moving targets. Since each image captured by such a sensor is a spatio-temporal image, where different parts of a moving object are viewed at different times and different viewpoints, it is interesting to study the 3D and motion estimation problems with a stereo vision system with such imaging geometry. We present our dynamic pushbroom stereo
vision model under a linear camera motion model and piece-wise object motion models. We have shown that such a stereo system has uniform depth resolution. We also provide methods to extract both 3D and motion information from a dynamic pushbroom stereo image pair.

Based on the proposed models, we have studied two examples for surveillance and inspection. In the first example, we present a practical approach for 3D measurements in gamma-ray (or X-ray) cargo inspection. Thanks to the constraints of the real scanning system, we model the system by using a linear pushbroom sensor model with only one rotation angle instead of three. This greatly simplifies the calibration procedure and increases the robustness of the parameter estimation. Using only the knowledge of the dimensions of the cargo container, we can automatically calibrate the sensor and find all the sensor parameters, including the image center, the focal length, the 3D sensor starting location, the viewing direction, and the scanning speed. The sensor modeling and calibration is accurate enough for 3D measurements. Then, a semi-automated stereo reconstruction approach is proposed to obtain 3D measurements of objects inside the cargo. With both the interactive matching procedure and the 3D visualization interface, the 3D measurements for cargo inspection could be put into practical use.

In the second example, we present a new approach to extract both 3D structure and independent moving targets from long video sequences captured by an airborne camera. The principle of dynamic pushbroom stereo mosaics is presented. Based on the properties of the dynamic pushbroom stereo, we propose a new segmentation-based stereo matching approach for both 3D reconstruction and moving target extraction from a pair of dynamic pushbroom stereo mosaics for urban scenes. A simple yet effective natural matching primitive selection method is provided. This method is effective for stereo matching of man-made scenes, particularly when both 3D facilities and moving targets need to be extracted. We discussed the natural-primitive-based matching approach in the scenario of parallel-perspective pushbroom stereo geometry, but apparently the method is also applicable to other types of stereo geometry such as perspective stereo, full parallel stereo, and circular projection panoramic stereo.

The dynamic linear pushbroom stereo model provides a new imaging geometry that can find applications in widely used scanning systems for surveillance and inspection, including X-ray, gamma-ray, visible video, and IR video where the sensors are moving. This model unifies both the advantages of uniform depth resolution of the pushbroom stereo and the capability of independent moving target detection of the extended model to a dynamic version.
We have shown some early but promising results for two real applications. Several research issues need to be further investigated. First, we only provide methods to infer 3D and motion for dynamic objects using some physical constraints (Section 2.3). More general methods are needed to infer both 3D and independent motion using the dynamic pushbroom stereo. Second, little work exists in performing stereo matching on gamma-ray or X-ray images. We have not applied the dynamic model in this type of imaging system (Section 3). Automated stereo matching and moving target extraction methods for X-ray /gamma-ray pushbroom stereo will be a further research issue. Knowledge of physics and optics in generating the radiographic images could be very helpful in advancing this direction of research. Third, the experiments we performed for 3D and moving target extraction from aerial video (Section 4) are based on the model of generalized stereo mosaics we have previously proposed (Zhu, et al, 2004), in which the viewpoints are along a 3D curve instead of a more desirable straight line as in the linear pushbroom imaging model we proposed in this text. Automated generation of truly linear pushbroom stereo mosaics from a video sequence requires accurate camera orientation estimation of many frames of a lone video sequence and full ray interpolation to generate true linear pushbroom geometry. This will allow an analysis of the accuracy of the 3D and motion recovery using the dynamic pushbroom stereo model. All of these are our on-going efforts.

ACKNOWLEDGEMENTS

This work is supported by the Air Force Research Laboratory (AFRL) under the RASER Program (Award No. FA8650-05-1-1853), and is partially supported by an AFRL Human Effectiveness Program (Award No. F33615-03-1-63-83), by Army Research Office (ARO) under Award No. W911NF-05-1-0011, and by the CUNY Graduate Research Technology Initiative (GRTI) program. We thank Rex Richardson and Victor J. Orphan at Science Applications International Corporation (SAIC) for providing us gamma-ray images and the dimension data of the cargo container. The aerial video data were captured in the UMass Computer Vision Lab under a NSF grant (No EIA-9726401). Thanks are also given to Hao Tang, Bing Shen, Li Zhao and Jiaye Lei for implementing the algorithms of the two applications, and to Robert Hill for proofreading the manuscript, all at the City College of New York. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright notation thereon. However, the views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the
official policies or endorsements, either expressed or implied, of the U.S. Government.

REFERENCES


Tao, H., H. S. Sawhney and R. Kumar, 2001. A global matching framework for stereo computation, In *Proc. ICCV’01*


Xiao, J. and M. Shah, 2004. Motion layer extraction in the presence of occlusion using graph cut, In *Proc. CVPR’04*


