The Third International Symposium on 3D Data Processing, Visualization and Transmission (3DPVT 2006), University of North Carolina, Chapel Hill, USA, June 14-16, 2006

Dynamic 3D Urban Scene Modeling Using Multiple Pushbroom Mosaics

Hao Tang, Zhigang Zhu, George Wolberg
Department of Computer Science, The City College of New York, New York, NY 10031
{tang | zhu | wolberg}@cs.ccny.cuny.edu

Abstract

In this paper, a unified, segmentation-based approach is proposed to deal with both stereo reconstruction and moving objects detection problems using multiple stereo mosaics. Each set of parallel-perspective (pushbroom) stereo mosaics is generated from a video sequence captured by a single video camera. First a color-segmentation approach is used to extract the so-called natural matching primitives from a reference view of a pair of stereo mosaics to facilitate both 3D reconstruction of textureless urban scenes and man-made moving targets (e.g. vehicles). Multiple pairs of stereo mosaics are used to improve the accuracy and robustness in 3D recovery and occlusion handling. Moving targets are detected by inspecting their 3D anomalies, either violating the epipolar geometry of the pushbroom stereo or exhibiting abnormal 3D structure. Experimental results on both simulated and real video sequences are provided to show the effectiveness of our approach.

1. Introduction

Mosaics have become common for representing a set of images gathered by one or more (moving) cameras. We are particularly interested in parallel-perspective mosaics with *pushbroom stereo* geometry [1, 24, 26]. The term "pushbroom" is borrowed from satellite pushbroom imaging [8] where a linear pushbroom camera is used. The basic idea of the pushbroom stereo mosaics is as follows. If we assume the motion of a camera is a 1D translation and the optical axis is perpendicular to the motion, then we can generate two spatio-temporal images (mosaics) by extracting two scanlines of pixels of each frame, one in the leading edge and the other in the trailing edge. Each mosaic image thus generated is similar to a *parallel-perspective* image captured by a linear pushbroom camera [8], which has parallel projection in the direction of the camera's motion and perspective projection in the direction perpendicular to that motion. Pushbroom stereo mosaics have uniform depth resolution, which is better than with perspective stereo, or the multi-perspective stereo with circular projection [16, 18]. Pushbroom stereo mosaics can be used in applications where the motion of the camera has a dominant translational direction. Examples include satellite pushbroom imaging [8], airborne video surveillance[24, 26], 3D reconstruction for image-based rendering [1], road scene representations for robot navigation [22, 25], under-vehicle inspection [5, 11], and 3D measurements of industrial parts by an X-ray scanning system [8, 14], and of articles in gamma-ray cargo inspection [27]. 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 [2, 21, 23].

Stereo vision is one of the most important topics in computer vision, and recently a through comparison study [6] has been performed. Simple window-based correlation approaches do not work well for man-made scenes. In the past, an adaptive window approach [9] and a nine-window approach [7] are used to deal with some of these issues. Recently, color segmentation has been used for refining an initial depth map to get sharp depth boundaries and to obtain depth values for textureless areas [20], and for accurate layer extraction [10]. Stereo matching algorithms using energy minimization frameworks [e.g., 4 and 19] can obtain accurate depth information; however, in addition to retrieving the accurate depth information, detecting moving objects and obtaining higher level object representations are also our goals. An interactive method has been proposed to obtain accurate object modeling using lines and structures [12], but we want to develop fully automatic algorithms that work for both man-made and natural scenes.

In this paper, 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 3D reconstruction and moving target extraction from 3D urban scenes. But the idea is applicable to more general scenes and other types of stereo geometry.

The paper is organized as follows. In Section 2, the mathematical framework of the dynamic pushbroom stereo is given, and then its properties for moving target extraction are discussed. In Section 3, multi-view pushbroom mosaics are proposed for extracting 3D structure and moving targets. In Section 4, our stereo matching algorithm for 3D static and moving target extraction will be provided. Experimental results will be given in Section 5 with both simulated and real video data. Finally is a brief summary in Section 6.

2. Dynamic Pushbroom Stereo Mosaics

Dynamic pushbroom stereo mosaics are generated in the same way as with the static pushbroom stereo mosaics described above. Fig.1 illustrates the geometry. A 3D point $P(X,Y,Z)$ on a target is first seen through the leading edge of an image frame when the camera is at location L_1. If the point P is static, we can expect to see it through the trailing edge of an image frame when the camera is at location L_2. The distance between leading and trailing edges is d_y (pixels), which denotes the constant "disparity". However, if point P moves during that time, the camera needs to be at a different location L'_2 to see this moving point through its trailing edge. For simplifying equations, we assume that the motion of the moving points between two observations (L_1 and L'_2) is a 2D motion (S_x, S_y), which indicates that the depth of the point does not change over that period of time. Therefore, the "depth" of the moving point can be calculated as

$$Z = F\frac{B_y - S_y}{d_y} \qquad (1)$$

where F is the focal length of the camera and B_y is the distance of the two camera locations (in the y direction). Mapping this relation into stereo mosaics following the notation our previous work [26], we have

$$Z = H\left(\frac{d_y + \Delta y - s_y}{d_y}\right) \qquad (2)$$

and $$(S_x, S_y) = (Z\frac{s_x}{F}, H\frac{s_y}{F}) = (Z\frac{\Delta x}{F}, H\frac{s_y}{F}) \qquad (3)$$

where H is the depth of plane on which we want to align our stereo mosaics, (Δx, Δy) is visual motion in the stereo mosaics of the moving 3D point P, and (s_x , s_y) is the target motion represented in stereo mosaics. Obviously, we have $s_x = \Delta x$.

We have the following interesting observations about the dynamic pushbroom stereo geometry for moving target extraction. (1) *Stereo fixation*. For a static point (i.e. $S_x = S_y = 0$), the visual motion of the point with a depth H is (0,0), indicating that the stereo mosaics thus generated fixate on the plane of depth H. (2) *Motion accumulation*. For a moving point ($S_x \neq 0$ and/or $S_y \neq 0$), the motion between two observations accumulates over a period of time due to the large distance between the leading and trailing edges. (3) *Epipolar constraints.* In the ideal case of 1D translation of the camera (with which we present our dynamic pushbroom stereo geometry), the correspondences of static points are along horizontal epipolar lines, i.e. $\Delta x = 0$. Therefore, for a moving target P, the visual motion with nonzero Δx will identify itself from the static background in the general case when the motion of the target in the x direction is not zero (i.e., $S_x \neq 0$). (4) *3D constraints.* Even if the motion of the target happens to be in the direction of the camera's motion (i.e. the y direction), we can still discriminate the moving target by examining 3D anomalies. Typically, a moving target (a vehicle or a human) moves on the flat ground surface (i.e., road) over the time period when it is observed through the two edges of the video images. A moving target in the direction of camera movement, when treated as a static target, will show 3D anomaly - either hanging up above the road (when it moves to the opposite direction, i.e., $S_y < 0$), or hiding below the road (when it moves in the same direction, i.e., $S_y > 0$).

After a moving target has been identified, the motion parameters of the moving target can be estimated. We first estimate the depth of its surroundings and apply this depth Z to the target, then calculate the object motion s_y using Eq. (2) and (S_x, S_y), using Eq. (3), given the known visual motion (Δx,Δy) observed in the stereo mosaics.

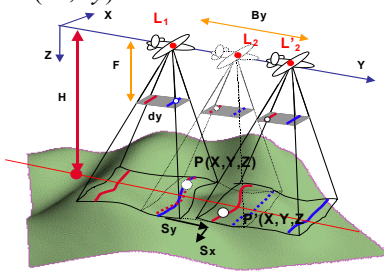

Fig. 1. Dynamic pushbroom stereo mosaics

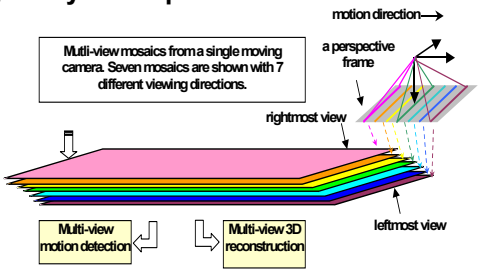

Fig. 2. Multi-view pushbroom mosaics

3. Multi-View Pushbroom Mosaics

A pair of stereo mosaics (generated from the leading and trailing edges) is a very efficient representation for

both 3D structures and target movements. However, stereo matching will be difficult due to the largely separated parallel views of the stereo pair. Therefore we propose to generate multi-view mosaics (more than 2), each of them with a set of parallel rays whose viewing direction is between the leading and the trailing edges (Fig. 2). The multiple mosaic representation is still efficient. Moreover, there are some benefits of using them. First, it eases the stereo correspondence problem in the same way as the multi-baseline stereo [15], particularly for more accurate 3D estimation and occlusion handling. Second, multiple mosaics also increase the possibility to detect moving targets with unusual movements and also to distinguish the movements of the specified targets (e.g., ground vehicles) from those of trees or flags in wind. In the next section, we will discuss a new method to extract both of the 3D buildings and moving targets from the stereo mosaics.

4. Dynamic and 3D Content Extraction

Using the advantageous properties of multi-view mosaics, we propose a unified approach to perform both stereo matching and motion detection. In a set of pushbroom mosaics generated from a video sequence, the leftmost mosaic is used as the reference mosaic, therefore color segmentation is performed on this mosaic, and the so called *natural matching primitives* are extracted. Multiple natural matching primitives are defined with each homogeneous color image patch approximately corresponding to a planar patch in 3D. The representations are effective for both static and moving targets in man-made urban scenes with objects of largely textureless regions and sharp depth boundaries. Then matches of those natural matching primitives are searched in the rest of the mosaics, one by one. After matching each stereo pair, a plane is fitted for each patch, and a set of planar parameters for the planar patch is estimated. Then multi-view matches are performed, and therefore multiple sets of parametric estimates for this planar patch are obtained. The best set is selected as the final result by comparing match evaluation scores. In the following subsections, we will describe the approach in more detail.

4.1. Patch and interest point extraction

First, the reference mosaic of the stereo mosaic pair, is segmented, using the mean-shift-based approach [3]. The segmented image consists of image regions (patches) with homogeneous color, and each of them is assumed to be a planar region in 3D space. For each patch, the boundary is defined as a closed curve. Then we use a line fitting approach to extract feature points for stereo matching. The boundary of each patch is first fitted with connected straight-line segments using an iterative curve splitting method. The connecting points between line segments are defined as *interest points*, around which the natural matching primitives are going to be defined. Now we are ready to perform the three-step stereo match.

4.2. Three-step stereo match

A three-step matching algorithm is utilized on the first stereo mosaic pair. Let the leftmost mosaic and the second mosaic be denoted as I_1 and I_2, respectively. The matching process consists of the following three steps.

Step 1: *Global match*. In a typical urban scene, for a frontal or near-frontal surface, all the pixels inside the patch (region) have similar visual displacements. Therefore, for each region in the mosaic I_1, the sum of absolute difference (SAD) is carried out for all pixels in this region between the two mosaics I_1 and I_2 with a preset search range, defined as

$$SAD(\Delta y) = \sum_{(x,y)\in R} | I_1(x,y) - I_2(x + \Delta x(y + \Delta y), y + \Delta y) | \qquad (4)$$

where the summation is carried out for all the pixels inside the region (patch) **R**, and Δy is the visual displacement in the y direction. The function $\Delta x(.)$ denotes the epipolar function (it is zero for the ideal case) [26]. Thus the initial visual displacement of the region between I_1 and I_2 is obtained as the one minimizing the SAD.

Step 2: *Local match*. Since not all regions are frontal planes in 3D space, the pixels in each region do not have a fixed visual displacement. Thus, for each interest point, the best match is searched within a neighborhood area of the initial visual displacement. Instead of using the conventional window-based match, we define the so-called *natural matching primitives* (Fig. 3) to conduct a sub-pixel stereo match. We define a region mask M of size mxm centered at that interest point such that

$$M(i,j) = \begin{cases} 1, \text{ if } (x + i, y + j) \in \mathbf{R} \\ 0, \text{ otherwise} \end{cases} \qquad (5)$$

The size m of the natural window is adaptively changed depending on the size of the region **R**. In order that a few more pixels (1-2) around the region boundary (but not belonging to the region) are also included so that we have sufficient image features to match, a dilation operation is applied to the mask M to generate a region mask covering pixels across the depth boundary. The weighted cross-correlation, based on the natural window centered at the point *(x, y)* in the reference mosaic, is defined as

$$C(\Delta x, \Delta y) = \frac{\sum_{i,j} M(i,j) I_1(x+i, y+j) I_2(x+i+\Delta x, y+j+\Delta y)}{\sum_{i,j} M(i,j)} \qquad (6)$$

Note that we still carry out correlation between two color images but only on those interest points on each region

boundary, and only with those pixels within the region and on the boundaries. A sub-pixel search is performed in order to improve the accuracy of 3D reconstruction; and a match is marked as reliable if it passes the crosscheck [6]

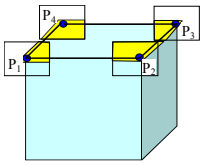

Fig. 3. Natural matching primitives

Step 3: *Surface fitting*. Assuming that each homogeneous color region is planar in 3D, a 3D plane $aX+bY+cZ=d$, which is represented in the camera coordinate system as shown in Fig. 1, is fitted to each region after obtaining the 3D coordinates of the interest points of the region using the pushbroom stereo geometry (Eq. 2). We use a RANSAC method [13] to fit plane.

4.3. Plane parameters from multiple mosaics

After the above three steps are applied to the first pair of stereo mosaics, initial estimations of the 3D structure of all the patches (regions) in the reference mosaic are obtained. Further matches between the reference mosaic and each of the rest of the mosaics are then conducted. However the global match (Step 1 in the section 4.2) is not applied; instead, the initial visual displacement of each interest point on a patch is predicted from the result of this point estimated from the first stereo pair. From Eq. (2), we know the visual displacement Δy is proportional to the selected "disparity" (d_y) for a pair of stereo mosaics for any static point (i.e., $s_y = 0$). Therefore the visual displacement of the interest point in consideration can be predicted except when the point is on a moving object that (1) does not move along the epipolar line of the pushbroom stereo; or (2) moves along the epipolar line, but with a varying speed. Those points will be reconsidered in the moving target detection stage. For refining the initial estimates of visual displacements based on the predictions from the results of the first pair of stereo mosaics, Steps 2 and 3 in Section 4.2 are performed to obtain new plane parameters for each pair of stereo mosaics starting from the second pair.

Suppose there are N pairs of stereo mosaics, constructed from N+1 pushbroom mosaics. Then N sets of plane parameters (a_k, b_k, c_k, d_k), $k=1,\ldots,N$, are obtained for each region (patch) in the reference mosaic. In order to obtain the most accurate plane parameters for each planar patch, the following steps are performed. First, for each pair of stereo mosaics, the patches in the reference mosaic are warped to the target mosaic in order to compute a color sum of absolute differences (SAD) for each region, between warped and original target images. Then, among all the estimates for each patch, the set of plane parameters with the least SAD value is selected as the best plane estimate. Note that using the knowledge of plane structure (i.e., 3D orientation), the best angle to view the region can be estimated, where the viewing direction of the selected mosaic (among all the possible viewing directions) is as close as to the plane norm direction. For example, for the side of a building that faces the left in Fig. 2, the best match could be obtained from first pair of stereo mosaics. If the view angle is equal to or greater than 90 degrees (relative to the plane norm), the region will not be visible. Incorporating this information, the SAD calculations are only carried out for those patches between the reference and target mosaics if the plane norms have less than 90-degree view angles from the viewing directions of the mosaics.

4.4. Plane merging

After the plane parameters with the smallest SAD value have been obtained for each region, we will have a close look at the best SAD of each region. If the SAD value is less than a preset threshold, then the patch is marked as *reliable*. We have found that a large number of small regions around a large region corresponding to a surface (or part) of a 3D object are generated by color segmentation, and they are difficult to obtain accurate 3D estimates because of the lack of sufficient feature points. Therefore, we perform a modified version of the neighboring plane parameter hypothesis approach [20] to infer better plane estimates. The main modification is that the parameters of a neighboring region are adopted only if it is marked reliable and the best neighboring plane parameters are accepted only when the match evaluation cost using the parameters is less than a threshold. The neighboring regions sharing the same plane parameter are then merged into one reliable region. This step is performed recursively till no more merges occur. We prefer to have false negatives than false positives, and the former will be handled in the next stage – moving object detection, which is our second major goal.

4.5. Moving object detection

After the plane merging stage, most of the small regions are merged together and marked as reliable. Moving object patches that move along epipolar lines should obtain reliable matches after the plane merging step, but they appear to be "floating" in air on below the surrounding ground, with depth discontinuities all around it. In other words, they can be identified by checking their 3D anomalies.

In general cases, most of the moving targets are not exactly on the direction of the camera's motion, those

regions should have been marked as unreliable in the previous steps. Regions with unreliable matches fall into the following two categories: (1) moving objects with motion not obeying the pushbroom epipolar geometry; (2) occluded or partially occluded regions, or regions with large illumination changes. For regions in the second category, their SADs in stereo matching evaluation are always very high. The regions in the first category correspond to those moving objects that do not move in the direction of camera motion, therefore they do not obey the pushbroom stereo eipoloar geometry. Therefore, for each of these regions, we perform a 2D-range search within its neighborhood area, and a global match step similar to the first step of normal stereo matching (Sec. 4.2) is carried out for each such patch. If a good match (i.e., with a small SAD) is found within the 2D search range, then the region is marked as a *moving* object.

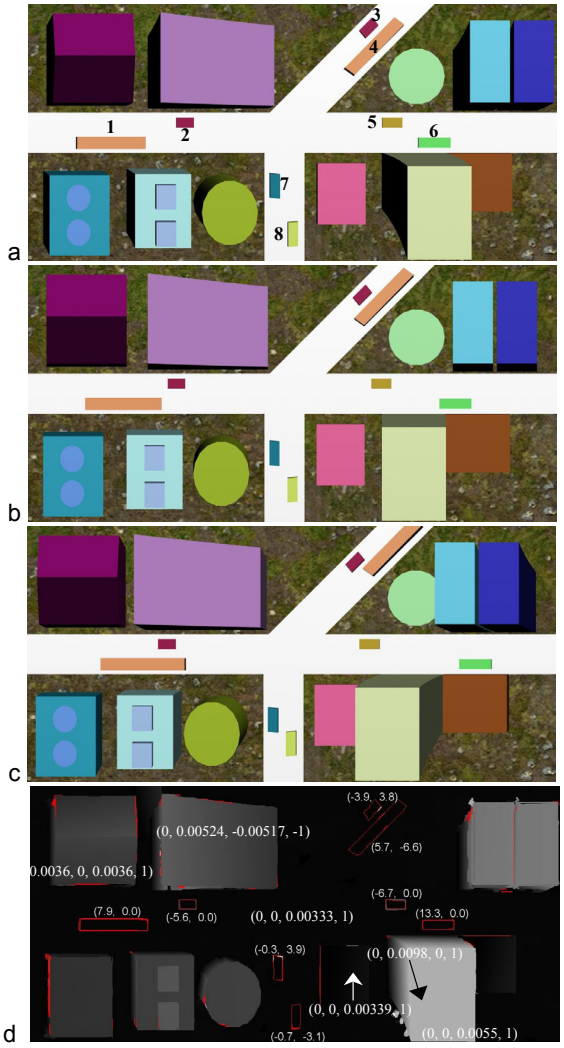

Fig 4. (a) The leftmost, (b) center and (c) rightmost views of the nine mosaics of a simulated scene. The final "height" map is shown in (d), labeled with plane parameters *(a,b,c,d)* for several representative surfaces (from left to right: one side of a ridged roof, a slanting roof, ground with depth Z= 300.0m, roof of a low building with Z = 295.0m, and side and roof of a tall building with Z=180.0 m), and motion displacements (s_x, s_y) of the detected moving targets.

5. Experimental Results and Analysis

Since it's difficult to obtain ground truth data for real video sequences, a simulated video sequence was generated with the ground truth data for the purpose of algorithm evaluation. Then, the proposed approach was applied to stereo mosaics generated from real world video sequences.

5.1 Results and analysis on a simulated scene

Nine parallel-perspective stereo mosaics were generated from a simulated video sequence of a simulated scene with ground truth data of both 3D and moving targets (Fig. 4). The virtual "aircraft" with a video camera flied at a 300-meter height above the scene along a 1D translational direction with a constant speed, and the motion direction is perpendicular to the optical axis of the camera. The 3D "buildings" are with heights from 5 to 120 meters above the ground, with different roof shapes.

Each of the eight moving objects undertakes a 2D translational motion with constant velocity during the period of the capture of the total 1700 frames of images, except the one labeled as "1" in Fig 4a. The velocity of the motion of each moving target is represented in centimeter (cm) per frame. Nine 1-column width slit windows are used to generate the nine mosaics (refer to Fig. 2), every pair of the two consecutive windows has a 40-pixel distance. Fig 4 shows three of the nine mosaics.

We compare the final estimated height map with the ground truth data. The error histogram (base 2 logarithmic scaling on the number of pixels) is shown in Fig. 5a for all the regions (including the moving object regions and other obvious wrong matches). From the error distribution, we have found that the errors of 86.5% points in the reference mosaic are within ±4 meters. The absolute average value of the errors for those points is only 0.317 meters. Note that in theory, the error of the depth/height estimation by the pushbroom stereo in Eq. 2 can be calculated as $\delta Z = (H/d_y)\,\delta y$, where δy is the error in stereo matching (in pixels). In this experiment, H is 300 meters, and d_y is from 40 to 320 pixels (from the first pair to the 8th pair of stereo mosaics), and ideally δy is 0.1 pixels with the sub-pixel local match step. Therefore, the theoretical errors after local match go from 0.75 down to about 0.1 meters from the first pair to the 8th pair.

However, larger viewing differences introduce larger errors in δy, therefore the error reduction by using larger disparities (from 40 to 320) is not as significant as the theoretical estimation. On the other hand, plane fitting on the multiple interest points with sub-pixel accuracy increases the accuracy in δZ, which leads to a more realistic error range close to the average error of the estimated depths/heights in this experiment (i.e., 0.317m). For showing how depth errors vary and how the planar parameters are selected among the eight pairs of stereo mosaics in generating the final height map, Fig. 5b shows the estimation errors of the planar parameters (from the ground truth) for the 17 largest regions in the reference mosaic. Most of the depth errors are below 0.3 meter, and the magnitudes are comparable among different pairs of stereo mosaics with various "disparities" (i.e., *dy*).

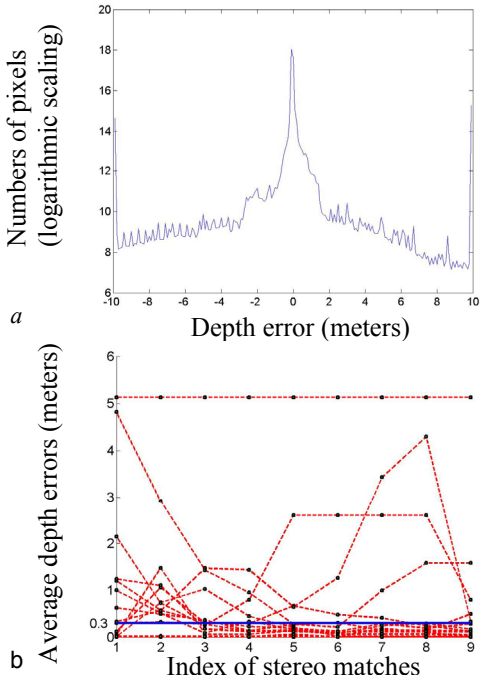

Fig 5. Depth error analysis. (a) Error histogram. (b) Comparison and selection among the results from the 8 pairs of stereo mosaics for the largest 17 regions. The last column (9th) shows the final selection.

After the regions have been merged, we analyze all the reliable regions, and those with obvious 3D anomalies are marked as moving objects (along the epipolar lines). For example, in Fig. 4.a., the heights of the regions labeled 1 and 6, if treated as static objects, are estimated as -39 meters and -50 meters high from the ground, respectively, much lower than the ground plane. The regions labeled 2 and 5 are estimated as 94 meters and 98 meters high from the ground, respectively, much higher than the ground. In fact all these regions only are 2 to 5 meters high from the ground. So these regions with such 3-D "anomalies" if incorrectly treated as static objects are detected as moving targets.

On the other hand, those unreliable regions (as possible candidates for moving objects not along the epipolar lines) further go through 2D-range searches for matches within their neighborhood areas (e.g., 30x30 2D range). In Fig. 4a, regions 3, 4, 7 and 8 are moving targets. They do not obtain reliable matches in the stereo match step, but could find reliable matches from their 2D range searches, between the first mosaic and the rest mosaics. Therefore they are considered as moving targets. Note that those regions marked with red boundaries in the height map have good matches in their 2-D range searches; however, many of them have very small sizes, or have very thin structures, therefore are not considered to be moving targets. The estimated motion parameters (s_x, s_y) of those detected moving targets from the first pair of stereo mosaics are marked in Fig. 4a (in pixels). The error analysis results of the 8 detected moving targets are shown in Table 1. The average error of the 2D motion estimation is (0.198, 0.008) in velocity (cm/frame), or (0.791, 0.033) in displacements (pixels) between the first pair of the stereo mosaics. The error for the 1st object is the largest since its velocity is not constant.

Table 1. Motion estimation errors

Obj Idx	Ground Truth (cm/frame)		Estimated Results (cm/frame)		Errors (cm/frame)	
	Sx	Sy	Sx*	Sy*	ΔSx	ΔSy
1	0	2.485	0	1.649	0	0.836
2	0	-1.499	0	-1.628	0	0.129
3	1.064	-1.262	1.053	-1.08	0.011	-0.181
4	-1.414	1.414	-1.444	1.247	0.031	0.166
5	0	-1.999	0	-2.012	0	0.013
6	0	2.499	0	2.495	0	0.003
7	0.999	0	0.982	-0.076	0.017	0.076
8	-0.781	0	-0.789	-0.178	0.007	0.178

5.2 Results on real video data

We also have performed experiments on pushbroom stereo mosaics from two real video sequences. The first group of mosaics (Fig. 6.) was generated from a ground video sequence of an indoor scene, the side view of several bookshelves and a file cabinet against a wall. Eleven mosaics were generated and used as input data for our algorithm to generate a height map of the entire scene. Three mosaics and the final "height" map are shown in Fig. 6, with the "height" values measured from the reference plane H. Note that height values are also obtained for textureless regions and thin structures.

The second group of mosaics was generated from an aerial video captured when the airplane was about 300 meters above the ground. Fig. 7a shows a pair of stereo mosaics from the nine mosaics generated from the video

sequence. Fig. 7b shows the close-up region of a window marked in Fig. 7a, which includes both various 3D structures and moving objects (vehicles). Fig.7c and 7d are "height" map generated using the proposed method. Note the sharp depth boundaries are obtained for the buildings with different heights and various roof shapes. The average heights of the buildings marked as No. 1 to No. 5 in Fig. 5c and Fig. 5d are 11.5m, 5.8m, 5.4m, 14.9m and 7.8m, respectively. The long building (No.1) has a slanting roof (left side is higher). Even though we have not conducted an accurate evaluation due to the lack of ground truth data, these estimations are consistent with the real data of these buildings. The moving objects that have been detected across all the nine mosaics are shown by their boundaries (in red). Those vehicles that are not detected by our algorithm are marked by rectangular bounding boxes; they are either stationary (as those in the boxes 2 and 3), or deformed differently across the mosaics due to the changes of motion in velocities (as in the box 1) and directions (as in the box 4).

6. Conclusions and Discussions

In this paper we present a new approach to extract both 3D structure and independent moving targets from long video sequences. The principles of dynamic pushbroom stereo mosaics are presented, which shows that the new geometry has advantages, in both moving object extraction and of 3D estimation, in terms of panoramic field of view, adaptive baseline system, independent motion accumulation, and parallel-perspective epipolar and 3D constraints for discriminating moving targets. The idea of a multi-view pushbroom mosaic is proposed to show the potential to estimate 3D structures of moving targets and to analyze different motion patterns.

Based on the properties of the dynamic pushbroom stereo mosaics, we propose a new segmentation-based stereo matching approach for both 3D reconstruction and moving target extraction from multi-view dynamic pushbroom stereo mosaics. 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 experimental results on simulated data show that the approach is both accurate in 3D reconstruction and effective in moving target detection. The preliminary experimental results on real video mosaics are also very promising. As our future work, three further studies are in consideration. First, in the current implementation, only 3D parametric information of planar patches in a reference mosaic is obtained. Since different visibilities are shown in mosaics with different viewing directions, we want to extend the approach presented in the paper to produce depth maps with multiple reference mosaics and then integrate the results by performing occlusion analysis. Second, the current implementation of moving object extraction and estimation is under the assumption that the velocity of each moving object is constant. Note that velocity changes of a moving object produce pushbroom image projections with changing sizes in mosaics. We want to make use of this piece of information to infer more complicated object motion parameters. Finally, we also want to conduct more experiments and evaluations on real video sequences of dynamic 3D urban scenes.

7. Acknowledgments

This work is supported by AFRL under Grant Award No. FA8650-05-1-1853, by NSF under Grant No. CNS-0551598, by ARO under Grant No. W911NF-05-1-0011, by ONR under Grant No. N000140310511 and by New York Institute for Advanced Studies.

8. References

[1] Chai, J. and H –Y. Shum, 2000. Parallel projections for stereo reconstruction. In Proc. CVPR'00: II 493-500.

[2] Collins, R., 2003. Mean-shift blob tracking through scale space, Proc.CVPR'03.

[3] Comanicu, D. and P. Meer, 2002. Mean shift: a robust approach toward feature space analysis. PAMI, May 2002

[4] Deng, Y., Q. Yang, X. Lin, and X. Tang. 2005 A symmetric patch-based correspondence model for occlusion handling. In Proc. ICCV'05 II: 1316-1322.

[5] Dickson, P., J. Li, Z. Zhu, A. Hanson,, E. Riseman, H. Sabrin, H. Schultz and G. Whitten, 2002. Mosaic generation for under-vehicle inspection. IEEE Workshop on Applications of Computer Vision Dec 3-4, 2002

[6] D. Scharstein and R. Szeliski. A Taxonomy and Evaluation of Dense Two-Frame Stereo Correspondence Algorithms. IJCV 47(1/2/3): 7-42, April-June 2002

[7] Fusiello, A., V. Roberto and E. Trucco, 1997. Efficient stereo with multiple windowing, In CVPR'97: 858-863

[8] Gupta R and Hartley R, 1997. Linear pushbroom cameras, IEEE Trans PAMI, 19(9), Sep. 1997: 963-975

[9] Kanade, T. and M. Okutomi, 1991. A Stereo Matching Algorithm with an Adaptive Window: Theory and Experiment, In Proc. IEEE ICRA'91, II: 1088-1095

[10] Ke, Q. and T. Kanade, 2001. A subspace approach to layer extraction, CVPR'01.

[11] Koschan, A., D. Page, J.-C. Ng, M. Abidi, D. Gorsich, and G. Gerhart, 2004. SAFER under vehicle inspection through video mosaic building," International Journal of Industrial Robot, Sep 2004, 31(5): 435-442

[12] Lee, S., Nevatia, R. Interactive 3D building modeling using a hierarchical representation, IEEE Workshop on Higher-Level Knowledge in 3D Modeling and Motion, Nice, Oct 2003

[13] Medioni, G., Kang, S. Emerging Topics in Computer Vision. Prentice Hall, ISBN: 0131013661 2004

[14] Noble, A., R. Hartley, J. Mundy and J. Farley. X-Ray Metrology for Quality Assurance, ICRA'94, II pp 1113-1119

[15] Okutomi M. and T. Kanade, 1993. A multiple-baseline stereo," IEEE Trans. PAMI, vol. 15, no. 4, pp. 353-363.

[16] Peleg, S., Ben-Ezra M and Pritch Y., 2001. Omnistereo: panoramic stereo imaging, IEEE Trans. PAMI, 23(3): 279-290

[17] Rousso, B., S. Peleg, I. Finci and A. Rav-Acha, 1998. Universal mosaicing using pipe projection, ICCV'98: 945-952.

[18] Shum, H.-Y. and Szeliski, R., 1999. Stereo reconstruction from multiperspective panoramas. In Proc. ICCV'99: 14-21

[19] Sun, J., Y. Li, S. Kang, and H-Y.Shum. Symmetric stereo matching for occlusion handling. CVPR'05. II: 399 - 406.

[20] Tao, H., H. S. Sawhney and R. Kumar, 2001. A global matching framework for stereo computation, ICCV'01

[21] Xiao, J. and M. Shah, 2004. Motion layer extraction in the presence of occlusion using graph cut, In Proc. CVPR'04

[22] Zheng, J. Y. and Tsuji, S. Panoramic Representation for route recognition by a mobile robot, Intl. J. Computer Vision, Vol. 9, no. 1, pp. 55-76, 1992.

[23] Zhou, Y. and H. Tao, 2003. A background layer model for object tracking through occlusion, ICCV'03: 1079-1085.

[24] Zhu, Z., E. M. Riseman and A. R. Hanson, 2001. Parallel-perspective stereo mosaics. In Proc. ICCV'01, vol I: 345-352.

[25] Zhu, Z. and A. R. Hanson, 2004. LAMP: 3D Layered, Adaptive-resolution and Multi-perspective Panorama - a New Scene Representation, Computer Vision and Image Understanding, 96(3), Dec 2004, pp 294-326.

[26] Zhu, Z., E. M. Riseman, A. R. Hanson, 2004. Generalized Parallel-Perspective Stereo Mosaics from Airborne Videos, PAMI, 26(2), Feb 2004, pp 226-237

[27] Zhu, Z., L. Zhao, J. Lei, 2005. 3D Measurements in Cargo Inspection with a Gamma-Ray Linear Pushbroom Stereo System, IEEE Workshop on Advanced 3D Imaging for Safety and Security, June 25, 2005, San Diego, CA, USA

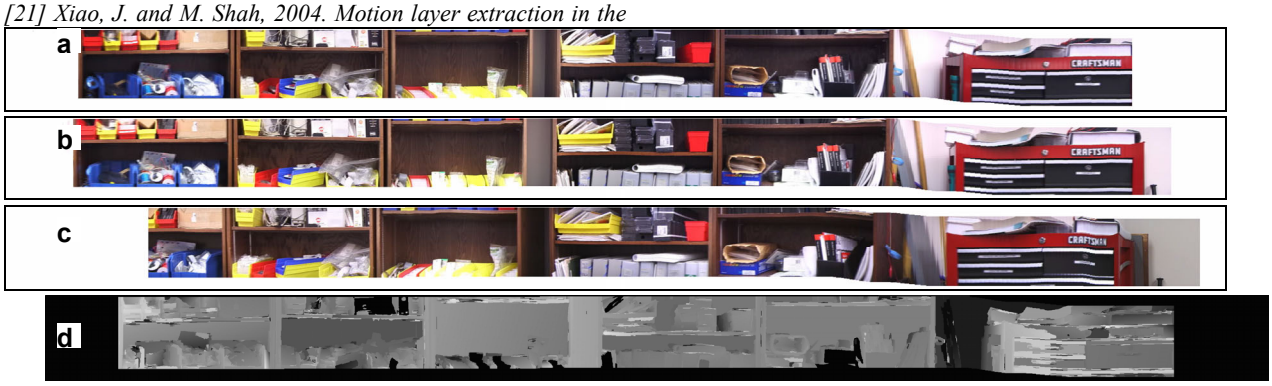

Fig 6. Multiview stereo mosaics for an indoor scene. (a) The leftmost, (b) center and (c) the rightmost views of total eleven mosaics. (d) the height map generated.

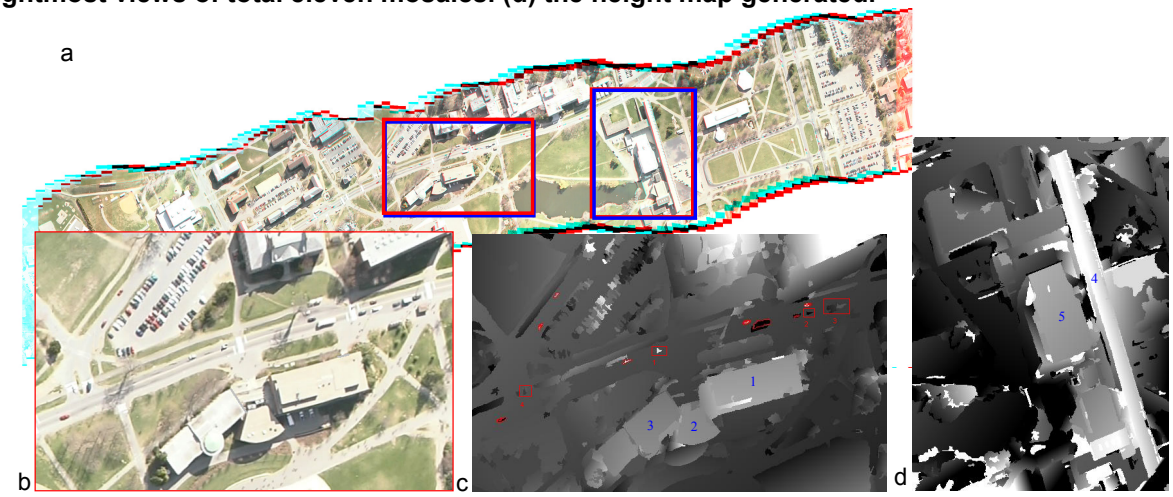

Fig 7. 3D and motion from multiview stereo mosaics of an aerial video sequence. (a) A pair of stereo mosaics from the total nine mosaics; (b) close-up of the 1st window marked in (a); and (c) the height map of the objects inside that window, with the detected moving targets marked by their boundaries and those not detected by rectangular boxes (see the electronic version); (d) height map of another part of mosaic (the 2nd window in (a)).