Exploiting Local and Global Scene Constraints in Modeling Large-Scale Dynamic 3D Scenes from Aerial Video

Hao Tang[†] Zhigang Zhu[§ †]

[§]Department of Computer Science, City College of New York, New York, NY 10031, USA
[†]Department of Computer Science, CUNY Graduate Center, New York, NY, 10016, USA
{tang, zhu}@cs.ccny.cuny.edu

Abstract

This paper presents a framework to analyze a large amount of video data and extract high-level structural information – planar structures and motion information - in typical urban scenes, which may be used in video coding or object recognition. The method consists of two phases. In the first phase, multiple parallel-perspective (pushbroom) mosaics are generated from the video data. In the second phase, the planar structures and the moving objects are extracted from the mosaics by a segmentation-based stereo match method.

The focus of this paper is the use of local and global scene constraints to improve the accuracy of high-level structural information extraction. The 3D planar patches obtained from the first step of 3D reconstruction are automatically clustered into one or more dominant planes, which are typical in indoor or outdoor city scene and are used to improve the 3D model. Then a local scene constraint is used to further refine the structure of a patch from the structures of its neighboring patches that have better structure estimations. Further, the dominant planes also provide information of road network directions, which greatly facilitates the search of moving objects on the roads. We demonstrate the effectiveness of our approach by experiments on a real video data set of a New York City scene.

1. Introduction

Extracting 3D and motion from video sequences of city scenes is very challenging for the following reasons. First, the 3D models of such scene change dramatically, e.g., with a lot of cluttered high-raising and low-raising buildings in a New York City scene we have been working. Second, the texture patterns are quite uneven – some regions are almost textureless while many have complicated surface structures. Third, these kinds of scenes typically have very shape depth boundaries and dealing with occlusions is an unavoidable issue. Fourth, independent moving targets are usually much smaller than the static structures, slower than the ego-motion of the sensor (particularly an airborne camera). Finally, lots of motion detection methods working on aerial video relying on first aligning the background and then detecting obvious intensity difference among aligned image sequences as independent moving target, but it's difficult to align the background in the city scene since there is large motion parallax in such cluttered aerial video. All these factors increase the difficulties in efficiently and accurately reconstructing 3D models and extracting moving targets from video sequences taken by a traveling camera.

To fulfill this goal, we want to generate dynamic "3D mosaics" from video sequences of city scenes taken by a traveling camera, with parallel-perspective *pushbroom* stereo geometry [2, 28]. Pushbroom stereo mosaics have uniform depth resolution that is better than with the perspective stereo, or the multi-perspective stereo with circular projection [19, 21].

In this paper, a set of parallel-perspective mosaics is generated to capture both 3D and dynamic aspects of the scene under the camera coverage. This step turns hundreds and thousands of frame images of a video sequence into just a few large field of view (FOV) mosaics. Though these large FOV mosaics are generated from a single camera, the results are much like using multiple line-scan cameras with different oblique angles (parallel viewing directions) to scan through the entire scene. Because of the multiple scanning angles, occluded regions in one mosaic can be seen from the others. Moving objects are all shown up in each mosaic, and by switching to different ones, the dynamic aspects can also be viewed and extracted.

Then a segmentation-based stereo algorithm designed is used to not only efficiently produces accurate matches across the depth boundaries, but also gives higher-level object structures since each object (e.g., a building) is represented into 3D planar regions and their relations. However, problems still remain dealing with many small color patches (due to over-segmentation) with unreliable matches and in detecting moving targets at directions different from the direction of ego-motion. Therefore, we explore the local and global constraints in facilitating 3D and motion extraction.

We note that many indoor and city scenes have one or more dominant planes, e.g., there are three mutual orthogonal plane directions in New York City scene, which can be a global constraint. In this paper, the local and global constraints are applied by the following method. First each homogenous patch, obtained by performing color segmentation on the reference mosaic and approximated as planar patch, is undertaken a stereo

matching process between the reference mosaic and a target mosaic. In this way, the plane parameters of all patches in the scene are computed. Second, an agglomerative clustering method is used to vote the dominant plane sets and then the information of the dominant planes is used to improve the accuracy of estimated structures. Third, a local scene constraint is used to further refine the structure of a patch from the structures of its neighboring patches that have better structure estimations. Fourth, moving objects are detected by checking 3D anomalies or applying 2D search in the target mosaic.

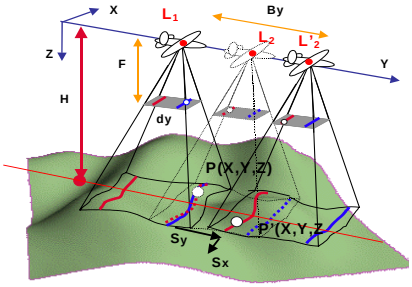

Fig. 1. Dynamic pushbroom stereo mosaics

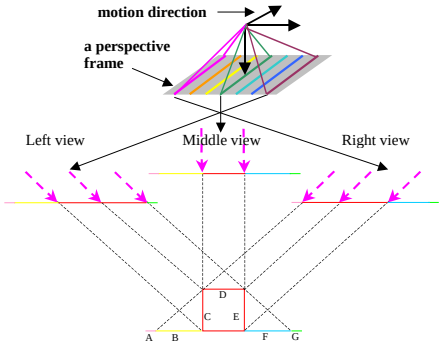

Fig. 2. Multi-view pushbroom mosaics

The rest of the paper is organized as follows. Section 2 provides a brief overview of some related work. Section 3 gives an overview of generation of pushbroom mosaics from a video sequence. Section 4 gives an overview of the segmentation-based multi-view stereo mosaic approach for 3D reconstruction. In Section 5, we discuss the extraction of dominant planes in the scene and applications of geometric constraints to improve 3D reconstruction results. Moving objects extraction is presented in Section 6. In Section 7 we show some experimental results and provide some concluding remarks in Section 8.

2. Related Work

Google Earth and Microsoft Virtual Earth map the earth by the superimposition of images obtained from satellite imagery, aerial photography and geographic information system data. Many buildings and structures from around the world now have detailed 3D structures, however, these 3D models are created manually and the number of city models is limited.

In order to reconstruct 3D models automatically, stereo vision is still one of the most important methods, and recently a thorough comparison study [20] has been performed. Global optimization based stereo matching methods, such as belief propagation [23] and graph cuts [1, 13], can obtain accurate depth information, but these methods are computationally expensive. Furthermore, in addition to retrieving the accurate depth information, detecting moving objects representations is also our goal.

Mosaics have become common for combining and representing a set of images gathered by one moving camera or multiple cameras. In the past, video mosaic approaches [10, 11, 14, 17] have been proposed for video representation and compression, but most of the work is for generating 2D mosaics instead of 3D panoramas, and using panning (rotating) cameras for arbitrary scenes or moving cameras for planar scenes, instead of traveling (translating) cameras typically used in airborne or ground mobile urban surveillance and 3D scene modeling. In the latter applications, obvious motion parallax is the main characterization of the video sequences due to the ego-motion of the sensors, obvious depth changes of the scenes, and independent moving targets. This paper deals with these problems with typical city or indoor scenes.

Combining intermediate depth images from several overlapping stereo pairs may not provide accurate results since accumulation error can quickly add up. Some work has been done in 3D reconstruction of panoramic mosaics [15, 22] with an off-center rotation camera, but the methods are limited to a fixed viewpoint camera instead of a moving camera; and the methods usually only deal with static scenes. For modeling large-scale 3D scenes, a new pushbroom sensor is developed in [9,16], which consists of nine CCD line sensors parallel to each other but with different viewing angles. Combining each line from nine sensors generates a pushbroom image, and semi-global matching method is used for 3D reconstruction. The geometric principle is the same as our stereo mosaics; however, we generate multiple mosaics from a single camera, and use a different 3D reconstruction approach. On the other hand, layered representations [12, 26, 27] have been studied for motion sequence representations; however, the methods are usually computationally expensive, and the outputs are typically motion segmentation represented by affine planes instead of true 3D information.

3. Multi-view Dynamic Stereo Mosaics

For completeness, we first give an introduction of the pushbroom stereo mosaics for a static scene. 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. Furthermore, dynamic pushbroom stereo mosaics are generated in the same way as 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

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

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 [28], we have

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

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

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 $s_x = \Delta x$. 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., $s_x \neq 0$). 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. While the geometry of pushbroom is presented with a pure 1D camera translation, the concept can be generalized to 6 DOF camera motion with a dominant motion direction.

A pair of stereo mosaics 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, multi-view mosaics (more than 2) are generated, each of them with a set of parallel rays whose viewing direction is between the leading and the trailing edges (Fig. 2). There are some benefits of using them. First, it eases the stereo correspondence problem in the same way as the multi-baseline stereo [18], particularly for improving accuracy of 3D estimation and handling occlusion. For example, object B is occluded in the right view while it appears in the left and middle views. The depth information of the patches B and D can be computed from a stereo match between the left view and the middle view mosaics. If the patches B and D in the reference mosaic (left view) are warped onto a target mosaic (e.g., right view) to be B' and D' under the known geometry of pushbroom mosaics, the warped patch B' is occluded by patch D'. Therefore, the match on patch B between left view mosaic and right view mosaic will not be correct. In another example, the structure (plane parameters) of the patch C can be computed from a stereo match between two mosaics (e.g., the left view and a view between the left and the middle views). Therefore, we know C is parallel to the middle view and cannot be observed from the middle view. So the match between reference mosaic (left view) and any views equal to or right to middle view will not be calculated. 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 three sections, we will discuss a method to extract both of the 3D buildings and moving targets from the stereo mosaics.

4. Multi-Mosaic Stereo Matching

In order to keep sharp depth boundaries and to obtain depth information for textureless areas, the reference mosaic (left view) is segmented into homogeneous color patches (regions) as the primitives for stereo matching. The interest points [24] with large curvature are extracted along the boundary of each homogenous patch. We use $Ip(l) = \{Ip_1, Ip_2, \ldots Ip_s\}$ denotes the set of interest points on *lth* patch, and s denotes the total number of the interest points in the patch.

Fig. 3. Outline of the stereo mosaic match algorithm

Modified Multi-view Stereo Match Algorithm

1. Perform color segmentation on the reference mosaic, and each homogenous color patch is approximated to be a planar surface.

2. The following three steps are applied

2.1. Extract interest points on the patch boundary;

2.2. Perform modified multi-scale matches on interest points of the region;

2.3. Use RANSAC to fit a plane for each homogenous patch

3. Refine plane parameters from multi-view mosaics

4. Each patch is classified as reliable or un-reliable by a match validation check

To carry out stereo matching, we use a modified version of our segmentation-based stereo match algorithm [24] that had a global matching step. An outline of the algorithm is shown in Fig. 3. Whereas the global match step assumes that the scene is frontal-parallel, the modified algorithm (without this step) is more general,

the modified algorithm is also more efficient since only a few interest points along the boundary of a patch (rather than all the points in the patch in the global match) are matched to obtain the 3D of all the points within the patch. Multiple pairs of stereo mosaics (refer to Fig. 2) are used for facilitating reliable stereo matching and more accurate 3D reconstruction. Suppose there are *N* pairs of stereo mosaics, constructed from *N+1* pushbroom mosaics. Then *N* sets of plane parameters $Pl(q)=(a_q, b_q, c_q, d_q)$, $q=1,\ldots,N$, are obtained for each patch in the reference mosaic. Only the best of N sets of plane parameters is selected as final result by a following warping and comparison process between the reference and other mosaics. If estimated plane parameters of a patch are correct, warping the patch in the reference image to the other views according to the plane parameters will render a patch that consistent with the real views. We formalize the above procedure using the following formulas. The match is calculated by following match cost function.

$$C(p, pl(q)) = \sum_{i=2}^{N+1} \sum_{(x,y)\in P} [(I_1(x,y) - I_i(x_i, y_i))^2] \qquad (4)$$

where *P* is the processing patch and *(x,y)* is one pixel in *P*. $(x_i, y_i) = f_i(p, pl(q))$ and $f_i(\bullet)$ is the function to calculate the correspondence (x_i, y_i) of a point *(x,y)* in the *ith* mosaic, $i=2,\ldots,N+1$. $I_i(\bullet)$ denotes the intensity of a point in the patch of the *ith* mosaic. The inner summation represents that of the square of intensity difference of a point in the patch on the reference mosaic and in the warped patch on the *ith* mosaic. The outer summation denotes that the inner computation process is performed on the reference mosaic with all other mosaics. Note: taking advantage of multi-view mosaics, if the patch is occluded in a mosaic, the match between reference mosaic and this mosaic is not taken into account. This is done by inferring the visibility of the patch in the *ith* view using the estimated plane parameters. The final parameter is selected by following equation.

$$P(k) = arg \min_k C(p, pl(k)) \qquad (5)$$

K represents the index of the set of plane parameters selected to be final result when a minimal summation is obtained.

Other modification of the algorithm is the use of a multi-scale local match (coarse to fine) approach with a match validation crosscheck:

$$| d_{(i,j)}(x,y) + d_{(j,i)}(x_j, y_j) | < h(l) \qquad (6)$$

where *(x,y)* and (x_j, y_j) are the correspondence pair (e.g., $(x,y)+d_{(i,j)}(x,y)=(x_j, y_j)$), $d_{(i,j)}(x,y)$ is a 2d disparity vector. The subscript *(i,j)* denotes that a match is performed from *ith* mosaic to *jth* mosaic. *h(l)* denotes a function of threshold in the *lth* step when the multi-scale match is performed and decreases from a large number to a fraction to obtain sub-pixel accuracy. Note, for each interest point, (1) it is located on the textured area (boundary of patch), and also with feature (large curvature). Hence it is robust on the match calculation; (2) a "natural" matching correlation template is used; only points in the considered regions are involved in the correlation computation (match cost step). Therefore, the correlation template is naturally adapted with the true object boundaries; and (3) a multi-scale approach is carried out, in that the search ranges and search steps are changed adaptively (from large to small) to achieve both robustness and efficiency. The multi-scale strategy is performed iteratively and usually converges in three steps. The region is marked reliable if the match cost is less than a threshold.

5. Refinement by Geometry Constraints

Dealing with 3D reconstruction of a large-scale scene from aerial video, lots of ambiguities and occlusions can degrade the accuracy, particularly for small patches after segment. Therefore, we explore two geometric constraints to further refine the estimation: local and global scene constraints.

5.1. Global scene constraint

We found that there are one or more dominant directions of planar surfaces in many indoor and outdoor city scenes. For example, for the outdoor city scenes, there exist three dominant planes (mutual perpendicular) in cities (e.g., New York City) and ground surface is a dominant plane in suburban scenes; an indoor scene obeying three mutual orthogonal planar surfaces are more popular due to the architectural structure. Therefore, some methods [4, 5] either rely on explicit edge detection and then find vanishing points, or use a direct Bayesian inference mechanism incorporating all of the image data to estimate three mutual orthogonal directions.

Thanks to the initial 3D reconstruction of our approach, some reliable matches and therefore structures of planar surfaces are obtained although many others may be still inaccurate. From reliable reconstructed planar surfaces, some dominant directions can be extracted. This geometric information can both benefit the improvement of 3D reconstruction and moving targets extraction. The detail procedures are explained as follows.

(1). Reliable matched patch can be classified if the match cost $C(\bullet)$ is less than a threshold.

(2). Collect all the norms of the planar patches with reliable match and calculate one or more dominant plane directions by an agglomerative clustering method [6].

(3). For the rest patches without reliable match, the dominant plane directions can be used to hypothesize initial estimates of these patches and then verified by calculating the match cost. If a smaller match cost is achieved for a patch then the hypothesis replaces the old unreliable result.

5.2 Local scene constraint

After applying the global scene constraint, 3D estimations are further refined by employing a local scene constraint. We perform a modified version of the neighboring plane parameter hypothesis approach [25] 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 cost (Eq. 4.) using the parameters is less than a threshold. Further, if the neighboring regions sharing the same plane parameter, then they 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.

6. Moving object detection

After performing geometric scene constraints, most of the small regions adopt the one of plane norms from the dominant directions or are merged with neighbors and marked as reliable. Moving object patches that move along epipolar lines should also obtain reliable matches after the plane merging step, but they appear to be "floating" in air or below the surrounding ground, with depth discontinuities all around it. In other words, they can be identified by checking their 3D anomalies. This is mostly true for aerial video sequences, where ground vehicles and humans move on the ground. Note that this is only the special case.

In general cases, most of the moving targets are not exactly on the direction of the camera's motion. Therefore, 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. 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 epipolar geometry. Therefore, for each of these regions, we can always perform a 2D-range search within its neighborhood area. If a good match (i.e., with a small Sum Squared Difference value) is found within the 2D search range, then the region is marked as a *moving* object. However, for regions in the second category, their 2D search in their neighborhood areas still cannot find any good match among all the mosaics. These regions are marked as occluded regions.

We can also take advantage of the known approximated road directions along which the traffic moves, to more effectively and more reliably search for matches of those moving vehicles. The road directions can be derived from 3D reconstruction results, e.g., in a city scene, the two dominant planes of the building façades surrounding the ground area on which the moving objects reside.

7. Experimental Results

In order to test the proposed framework, we have performed experiments on a real data set, New York City (NYC) scene. The NYC mosaics were generated from a video sequence from an NYC aerial video dataset. The video clip has about 758 frames of high-definition progressive video (1080*2000). Rooftops and city streets are seen as the camera looks ahead and down in a close flight just over One Penn Plaza and beyond in New York City. Yellow taxicabs make up a noticeable percentage of the vehicles traveling the grid of streets in this district of mostly lower-rising buildings, but with a few high-rise buildings (such as One Penn Plaza). Our main task is to recover the full 3D model of the area automatically, with cluttered buildings with various heights, from less than ten to more than a hundred meters. Fig. 5 shows one of the four multi-view mosaics (4816*2016) generated and used for 3D reconstruction and moving target detection. The camera moves from the left to the right in the mosaic.

This data set is very challenging due to the cluttered buildings and complex micro-surface structures that produce a lot of small homogeneous color patches after using mean-shift color segmentation [3]. In our experiments, among all of the regions that have successfully obtained plane-fitting results from multi-view mosaics, those with reliable matches (i.e., good match scores and reasonable sizes) are used to automatically vote for the three domination planes. The three plane sets supporting the dominant directions are plotted in Fig. 4. The three plane norms are [-0.549, -5.753, 1.000], [4.983, 1.493, 1.000] and [-0.209, 0.079, 1.000]. A simple cross-product check verifies they are approximately orthogonal to each other (The angles between them are 89.2^o, 109.5^o and 83.7^o). The information of these three domination plane directions is very useful in both refining the 3D reconstruction and extracting moving targets.

By making use of the global scene constraints, 3D reconstruction is refined by methods described in Section 5 namely using the three dominant planes and then neighborhood hypotheses. After these steps, the remaining regions, i.e., the "outliers", go through the moving object detection test. We use the method presented in Section 6. For this NYC dataset, we take advantage of the known road directions, to more effectively and more reliably search for matches of those moving vehicles.

Fig. 6a shows the 3D reconstruction results of the NYC video data, all represented in the reference mosaic. The height map is rendered from the result from the integration of the 3 stereo pairs of the mosaics (4 mosaics). Fig. 6b shows the colored coded height map (same as Fig. 6a). The color bar on the right-hand side shows the correspondences of colors and height values.

Due to the lack of the flight and camera parameters, we roughly estimate the main parameters of the camera from some known buildings. However, this gives us a good indication of how well we can obtain the 3D structure of this very complex scene. For example, the average heights of the three buildings at One Penn Plaza (marked as A, B and C in Fig. 6b) are 105.32 m, 48.83 m, and 19.93 m, respectively. Readers may visually check the heights of those buildings with Google Earth. Note that the building A has a low-rising part at the top- left corner, which was correctly reconstructed. Also note that the camera was not pointing perpendicularly down to the ground and therefore the reconstructed ground is tilted. This can be seen from the colors of the ground plane.

In order to test the performance of the global scene constraints on 3D reconstruction, we compare the match scores, the mean and standard deviation of match cost obtained from 3D reconstruction, are 1581.4 and 3338.6, respectively (without the global scene constraints), and 1447.2 and 3240.8, respectively (with global scene constraint). Note that the match scores are calculated on 3 pairs of mosaics, in three RGB channels. Although the statistics does not provide a large differences on the match scores due to that the unreliable matches only occur in small regions, the height maps is able to clearly show the improvement. Fig. 7 shows two examples. The two images on the left are the height maps of two small windows of the reference mosaic, obtained without using the global scene constraints. Quite some incorrect matches are shown in the height maps. The two images on the right are the height maps obtained by using the scene constraints. Apparently, most of errors are fixed.

The moving objects (vehicles) create "outliers" in the height map, as can be clearly seen on the color-coded map. For example, on the one-way road indicated in the first window (left) in Fig. 5, vehicles moved from the right to the left in the figure, therefore, their color-encoded height values have more red/yellow colors (i.e., the estimated heights are much higher than the ground if assumed static). On the other hand, on the one-way road indicated in the second window (right) in Fig. 5, vehicles moved from the left to the right in the figure, therefore, their color-encoded "height values have more blue colors (i.e., the estimated heights are much lower than the ground if assumed static). After further applying the information of the road directions obtained from the dominant plane voting, moving targets are efficiently and effectively searched and extracted. In Fig. 8, all of the *moving* targets (vehicles) are extracted, except the three circled in the figure. These three vehicles are merged with the road in color segmentation. Other vehicles that are not detected were stationary; most of them are on the orthogonal roads with red traffic signals on for stop, and a few parked on these two one-way roads.

8. Conclusions and Discussions

In this paper we propose a framework to both construct 3D model and detect moving target for long video sequences captured by a camera on a mobile platform. In the first step, multiple parallel-perspective (pushbroom) mosaics are generated to capture both the 3D and dynamic aspects of the scene under the camera coverage. In the second step, a multi-view, segmentation-based stereo matching approach is applied to extract parametric representation of the color, structure and motion of the dynamic and/or 3D objects.

The geometric constraints are applied to refine both 3D reconstruction and moving target detection. First, local scene supporting is used to refine 3D estimate. Second, an agglomerative clustering method is performed on the initial estimated planar structures to automatically vote the dominant plane directions which are used as good candidates of estimation of plane surfaces so that some global scene constraint can be used to fix some unreliable matches. Third, dominant planes are used to infer the grid of road directions in city scenes, which can be used to detect moving object more efficiently and robustly.

Three dominant plane directions of scenes like the New York City can greatly benefit 3D modeling and motion detection. The principle can be generalized to a more general urban or suburban scene, in which either single dominant ground plane and roofs, or more than three dominant planes can also be clustering to refine 3D models.

Modeling of large scale 3D scenes will have lots of important applications. In the future, our method can be extended to the following directions. First, because urban planning and developments are on a rapid pace, historical urban 3D scene models can be preserved by generating large scale mosaics and reconstructing 3D models. Further, changes over an urban scene can be detected by comparing two or more 3D models generated by different periods of times. Second, since a large scale urban scene is modeled by 3D planar structure by our approach and relations among all structures are known, we can recognize, label and index 3D structure (buildings) by further extracting high-level information from the 3D model.

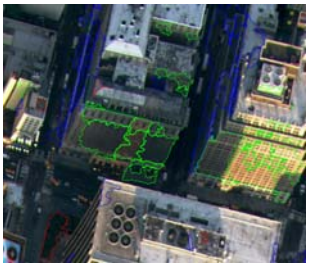

Fig. 4. Extracted dominant plane directions. Three mutual orthogonal plane directions represent norm directions of ground, front façades and side façades in red, green and blue, respectively. Only a small portion is shown here.

Fig. 5. A 4816 (W) x 2016 (H) mosaic from a 758-frame high-resolution NYC video sequence.

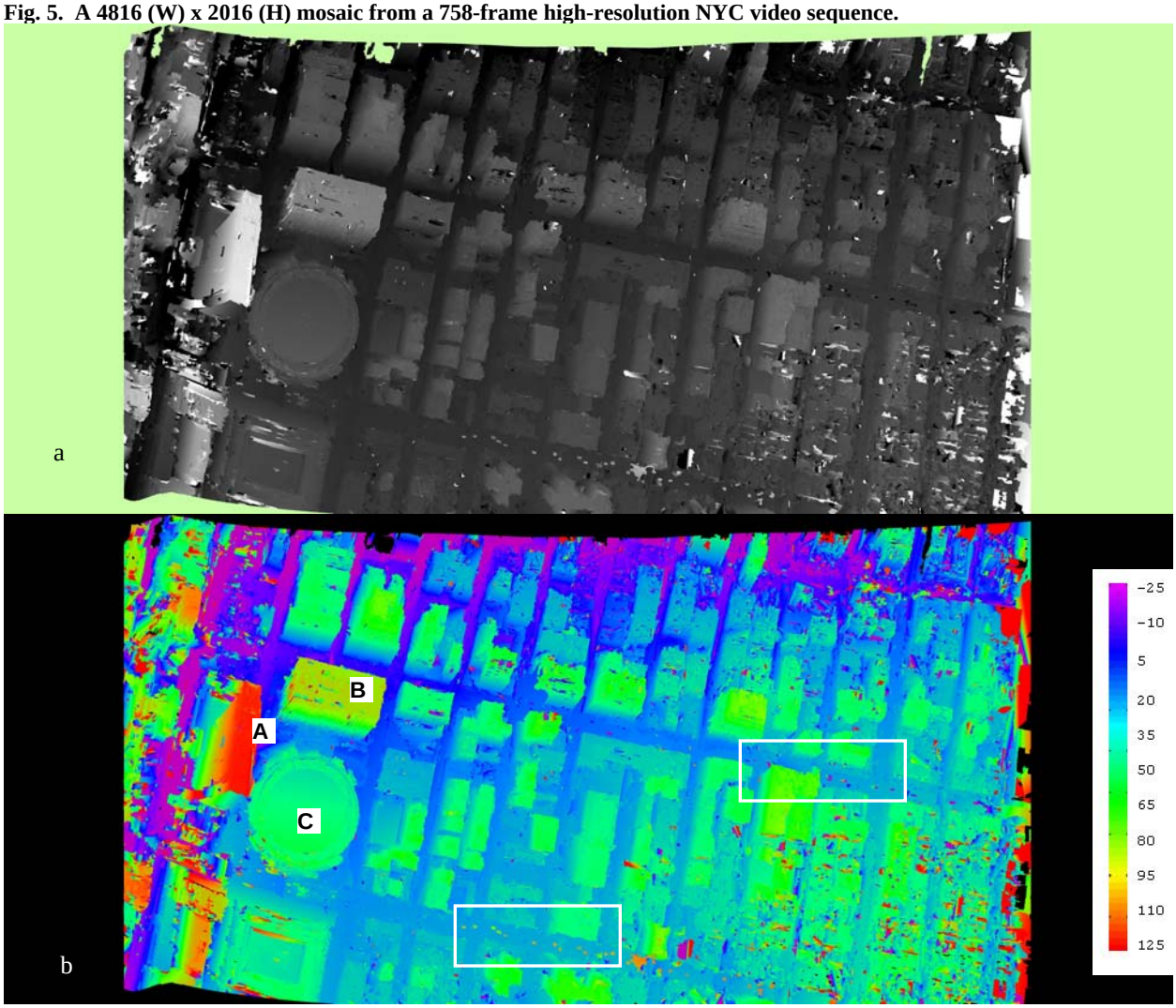

Fig. 6. (a) Height map from four mosaics, and (b) color-coded height map of (a)

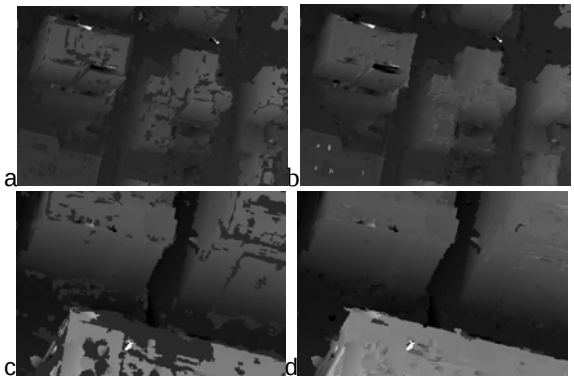

Fig. 7. Height maps of two small windows in the mosaic. a) and c) are height maps obtained without using the global scene constraints; and b) and d) are height maps obtained using the constraints.

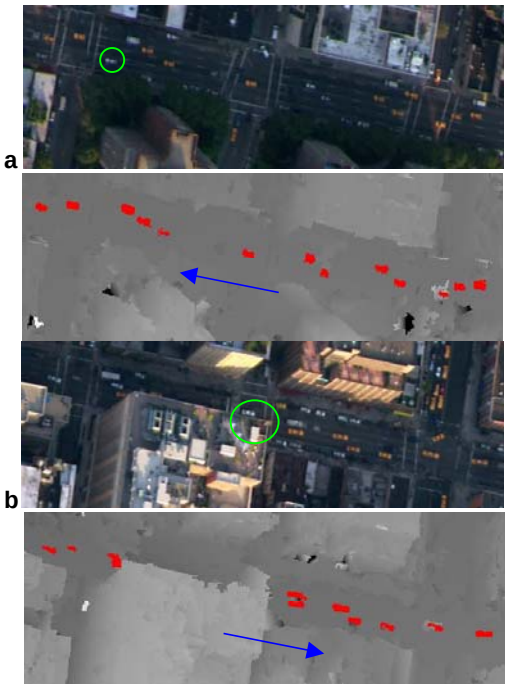

Fig. 8. Moving target detection using the road direction constraint. In the figure (a) and (b) are the corresponding mosaics and height maps of the down-left and the up-right windows in Fig. 5, with the detected moving targets painted in red. The two circles show the three moving targets that are not detected. The arrows indicate the directions of the roads along which the moving targets are searched.

Acknowledgements

This work is supported by AFRL/SN under Award No. FA8650-05-1-1853 and by NSF under Grant No. CNS-0551598.

References

[1] Boykov, Y., Veksler, O. and Zabih, R. Fast approximates energy minimization via graph cuts, PAMI, Vol. 23, No.11. 2001.

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

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

[4] Coughlan, C. and Yuille, A. 1999. Manhattan world: compass direction from a single image by Bayesian inference. ICCV, PP. 941-947.

[5] Deutscher, J.,Isard, M., MacCormick, J.P., Automatic camera calibration from a single manhattan image, ECCV02 (IV: 175).

[6] Dunham, M., Data Mining: Introductory and Advanced Topics, by, Prentice Hall, 2003

[7] 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

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

[9] Hirschmüller, H., Scholten, F. and Hirzinger, G. Stereo vision based reconstruction of huge urban areas from an airborne pushbroom camera (HRSC),DAGM-Symposium 2005, pp.58-66.

[10] Hsu, S. and Anandan, P., Hierarchical representations for mosaic based video compression, In Proc. Picture Coding Symp., 395-400, March 1996.

[11] Irani, M., Anandan, P., Bergen, J., Kumar, R. and Hsu, S.,. Mosaic representations of video sequences and their applications. Signal Processing: Image Communication, vol. 8, no. 4, May, 1996.

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

[13] Kolmogorov, V. and Zabih, R. Computing visual correspondence with occlusions using graph cuts, ICCV 2001.

[14] Leung, W. H. and Chen, T. Compression with mosaic prediction for image-based rendering applications, IEEE Intl. Conf. Multimedia & Expo., New York, July.

[15] Li, Y., Shum, H.-Y., Tang, C.-K. and Szeliski, R. Stereo reconstruction from multiperspective panoramas. IEEE Trans PAMI, 2004 26(1): pp 45-62.

[16] Neukum, G. The airborne HRSC-A, performance results and application potential, Pgotogrammetric Week'99, D. Frisch & R. Spiller, Eds., 1999.

[17] Odone, F., Fusiello, A. and Trucco, E. Robust motion segmentation for content-based video coding, the 6th Conference on Content-based Multimedia Information Access, College de France: 594-601. 2000.

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

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

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

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

[22] Sun, C. and Peleg, S. Fast Panoramic Stereo matching using cylindrical maximum surfaces, IEEE Trans. SMC Part B, 34, Feb. 2004. : 760-765.

[23] Sun, J. Zheng, N. and Shum, H. Stereo matching using belief propagation, PAMI, Vol.25, No. 7, July, 2003.

[24] Tang, H., Zhu, Z., Wolberg G. and Layne, J. R., 2006. Dynamic 3D urban scene modeling using Multiple pushbroom mosaics, the 3DPVT 2006, June 14-16, 2006.

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

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

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

[28] Zhu, Z., E. M. Riseman, A. R. Hanson, Generalized parallel-perspective stereo mosaics from airborne videos, PAMI, 26(2), Feb 2004