Stereovision-Based 3D Planar Surface Estimation for Wall-Climbing Robots

Hao Tang, *Student Member, IEEE*,
Zhigang Zhu, *Senior Member, IEEE*, Jizhong Xiao, *Senior Member, IEEE*

Abstract—**Finding traversable paths using computer vision is one of the most important components of an intelligent mobile robot system. For a wall climbing robot that operates in an urban environment, it is essential to automatically detect surface types and orientations for switching between moving and climbing, and for applying different adhesive forces both to save energy and ensure its own safety. This paper presents a novel segmentation-based stereovision approach in order to rapidly obtain accurate 3D estimations of urban scenes with largely textureless areas and sharp depth changes. The new approach takes advantage of the fact that many man-made objects in an urban setting consist of planar surfaces. Our approach has three main components: extraction of natural (planar) matching primitives, stereo matching via three-step algorithm (global match, local match and plane fitting), and plane merging and parameter refinement. Experimental results are provided for real indoor scenes.**

I. INTRODUCTION

In recent years, there have been strong demands to enlist robots for defense and counter-terrorism missions in urban environments. However, most of the mobile robots used in urban applications nowadays essentially move in 2-D planes without wall-climbing capability. It has been a long-time dream to develop miniature climbing robots with the ability to climb walls, walk on ceilings, and transit between different surfaces, thus transforming the present 2D world of mobile rovers into a new 3D universe. The robots we have designed and prototyped are named as City-Climbers (Fig 1) [3, 12], which overcome the limitations of the existing technologies in terms of climbing capability, re-configurability, and intelligence.

Robot navigation is a key component of an intelligent mobile robot system, especially of a wall climbing robot like the City-Climber. For a wall climbing robot that works in an urban environment, it is essential to automatically detect surface types and orientations for switching between moving and climbing, and for applying different adhesive forces both to save energy and ensure its own safety. In a typical indoor scenario, City-Climbers move on the ground, climb on walls and walk on ceilings. The geometric representations of the ground, walls and ceilings usually can be modeled as planes. In an outdoor setting, buildings and facilities that a climbing robot operates on usually consist of planar surfaces that are flat, e.g., facades and windows, or could be approximated locally as planar patches. Therefore, the problem can be mostly turn into a plane estimation and plane fitting problem.

This paper presents a novel segmentation-based stereovision approach in order to rapidly obtain accurate 3D estimations of urban scenes with largely textureless areas and sharp depth changes. Our approach has three main components: extraction of natural (planar) matching primitives, stereo matching via a three-step algorithm (global match, local match and plane fitting), and plane merging and parameter refinement. Experimental results are provided for real indoor scenes.

II. RELATED WORK

Various plane fitting methods have been proposed for robot navigation in both urban and terrain environments. Some of the methods use range finders to obtain 3D maps and then fit 3D planes on the range data. Weingarten, et al, [11] propose a generic method for plane fitting in an orthogonal least-square sense. By using laser range scanner scan, after filtering the outliers, matching planes are fused to find a compact 3D model of a scene. Ye and Borenstein [13] propose an obstacle negotiation method for mobile robots traversing rough terrain. They first transform a local terrain map surrounding the robot into a grid-type traversability map by extracting the slope and roughness of a terrain patch through least-squares plane-fitting; then reduce the amount of data in the 2-dimensional traversability map by transforming it into a 1-dimensional traversability field histogram, from which the velocity and the steering command are determined.

Other researchers use monocular vision systems to estimate ground plane. Liang and Pears [6] use the Kalman filter to track feature points (i.e., Harris corners) throughout an image sequence and group them into coplanar regions. By using this method, the ground plane patches could be extracted. Later, they extract planar homography relations of the ground plane from multi-view images [4]. Two point pair correspondences are used to compute the homography and to

This work is supported by ARO under Award No. W911NF-05-1-0011, AFRL under Award No. FA8650-05-1-1853, AFOSR under award No. FA9550-08-1-0199 and by NSF under Grant No. CNS-0551598 and IIS-0644127.

Hao Tang is with the Department of Computer Science, City College of New York, New York, NY 10031 (e-mail: tang@cs.ccny.cuny.edu).

Zhigang Zhu is with the Department of Computer Science, City College of New York, New York, NY 10031 (e-mail: zhu@cs.ccny.cuny.edu).

Jizhong Xiao is with the Department of Electrical Engineering, City College of New York, New York, NY 10031 (e-mail: jxiao@ccny.cuny.edu).

calculate heights of points above ground plane when camera undergoes a pure translation, and to extract camera and robot rotation from homographies under a general planar motion.

Researchers also use stereovision techniques to obtain depth maps and then to segment the depth maps for ground plane, curb or step estimation. Lu and Manduchi [5] present algorithms to detect and precisely localize curbs and stairways for autonomous navigation. They combine the edge information of images and 3-D data from a commercial stereo system to get the final results. Se, et al, [8] use a stereo vision system to match the scale-invariant feature transform (SIFT) visual landmarks, and robot ego-motion is estimated by least-squares minimization of the matched landmarks.

Various sophisticated stereo match algorithms have been developed and compared with test images [7]. Stereo matching algorithms using energy minimization frameworks [2, 9] could obtain fairly accurate depth maps, however, the computation involved is usually prohibitive for a real-time application, and the results are usually only depth information of point clouds, which have not provided meaningful structure information to be used for the wall climbing robot navigation.

III. City Climbers: Designs and Challenges

The adhesion device of the City-Climber is based on the aerodynamic attraction produced by a vacuum rotor package that generates a low-pressure zone enclosed by a chamber. The vacuum rotor package consists of a vacuum motor with impeller and exhaust cowling to direct air flow, as shown in Fig. 1a. It is essentially a radial flow device that combines two types of air flow. The high-speed rotation of the impeller causes the air to be accelerated toward the outer perimeter of the rotor, away from the center radically. Air is then pulled along the spin axis toward the device creating a low-pressure region, or partial vacuum region if sealed adequately, in front of the device. With the exhaust cowling, the resultant exhaust of air is directed toward the rear of the device, actually helping to increase the adhesion force by thrusting the device forward.

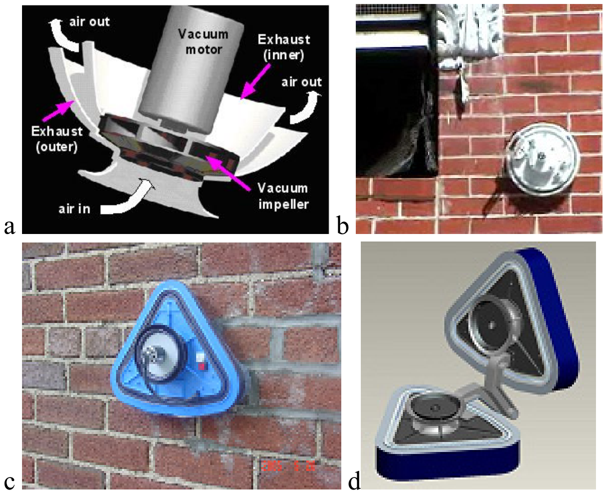

Fig. 1. Magnetization as a function of applied field. Note that “Fig.” is abbreviated. There is a period after the figure number, followed by two spaces. It is good practice to explain the significance of the figure in the caption.

Two low-pressure containment methods were investigated: inflated tube skirt seal and the flexible bristle skirt seal. To increase the mobility, a novel pressure force isolation rim is designed and the flexible bristle skirt seal made of re-foam is selected in current implementation. The driving system of an internal 3-wheeled drive and the payload are mounted on the platform, thus the re-foam makes the skirt and the robot system adaptable to the curve of rough surfaces. Technical details can be found in [3, 12]. Fig. 1 b) shows a City-Climber prototype-I operating on real brick wall via remote control.

The City-Climber prototype-II adopts the modular design which combines wheeled locomotion and articulated structure to achieve both quick motion of individual modules on planar surfaces and smooth wall-to-wall transition by a set of two modules. Each module can operate independently and is designed as triangle shape to reduce the torque needed by the hinge assembly to lift up the other module. To traverse between planar surfaces two climbing modules are operated in gang mode connected by a lift hinge assembly that positions one module relative to the other into three useful configurations: inline, +90°, and -90°. Responding the electronic controls, a sequence of translation and tilting actions can be executed that would result in the pair of modules navigating as a unit between two tangent planar surfaces; an example of this is going around a corner, or from a wall to the ceiling. Figure 1 c) shows the City-Climber prototype-II resting on a brick wall and Fig. 3 d) shows a conceptual drawing of two City-Climber modules operating in gang mode that allow the unit to make wall-to-wall and wall-to-ceiling transitions.

The City Climber robots are designed to move on various wall surfaces, such as brick, wood, glass, stucco, plaster, gypsum board, and metal. The video [3] illustrates the main areas of functionality and shows the results of several key experimental tests. The work in this paper is to make a City Climber autonomous and more intelligent by accurately estimate distances, orientations, areas and the relations of 3D planar surfaces in the view of the robot with a stereovision system. In our experiments, Bumblebee*,* a binocular stereo system, is used to capture stereo pairs.

Detection and localization of planar surfaces are important to a wall climbing robot for two reasons. First, the wall climbing robot need to safely and energy-efficiently stay on a surface (e.g., a wall or a ceiling). Surface orientation and roughness are useful information to adjust the suction motor speed to generate best adhesive force to the surface. Second, the wall climbing robot need to accurately transit between different surfaces (e.g. ground-to-wall, wall-to-ceiling, or wall-to-wall transitions). Therefore, knowing the distances and relations of surfaces are important. Unfortunately, for many navigation systems using simple stereovision techniques, only 3D point cloud information, which is also usually inaccurate, is offered. This cannot provide sufficient information for wall climbing robots to safely operate on walls and ceilings. High level parametric 3D structure representation could be extracted from accurate depth maps. However, in many sophisticated vision algorithms that can produce accurate depth maps, both stereo matching and structure analysis are time consuming procedures hence real time performance cannot be easily achieved. Moreover, large textureless or repetitive regions usually exist in an urban, man-made scenario, e.g. brick wall in Fig 1(c) and white walls and other flat surfaces as shown in Fig 2. Those regions bring in lots of ambiguities during the stereo matching step, therefore accurate 3D map cannot be easily obtained. In an indoor, textureless or weak textureless scene, lots of surfaces with uniform colors, hence correlation-based dense stereo match methods do not work well. Therefore, feature-based methods [4,6,8] have been widely used in robotic application. However, Feature-based methods can only provide sparse 3D data, leaving the task of planar surface estimation probably more difficult. Our approach takes advantages of both feature-based and dense stereo methods, so it can generate dense 3D geometry, and construct simple and useful 3D planar maps; but the matching is only performed on well-defined feature points so that it is both efficient and robust.

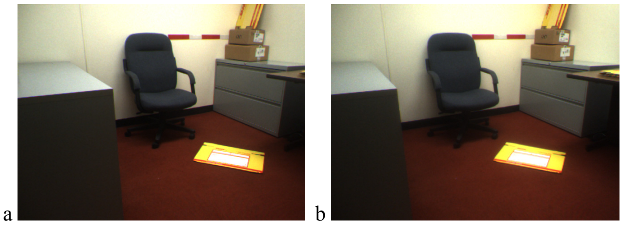

Fig 2. A pair of stereo images: (a) left image, and (b) right image.

IV. Our Approach

We propose a segmentation-based stereo vision approach to rapidly estimate 3D surface using plane patch matching and fitting. This approach has four advantages. First, depth information of textureless or weak-textured regions can be estimated. Second, accurate depth boundaries (i.e., 3D surface patch boundaries) can be accurately maintained. Third, 3D surface parameters, such as distances, orientations, boundaries and their relations are obtained, which can be directly applied to wall-climbing robot navigation. Finally, the computation is efficient since only selected feature points along boundaries of patches need to be matched. The algorithm currently runs at around 0.5 Hz for 320x240 images on the PIV 2.5GHz processor, therefore, real-time performance is possible with further code optimization and hardware acceleration.

Here is an overview of our approach. The left camera of binocular stereo system serves as the reference camera. First, color segmentation is performed on the reference image, and the so-called *natural matching primitives* are extracted. Multiple natural matching primitives are defined with each homogeneous color image patch corresponding to a planar patch in 3D space. The representations are effective for both outdoor and indoor scenes with objects of largely textureless regions and sharp depth boundaries. Then matches of those natural matching primitives are searched in the right image. After matching the stereo pair, a plane is fitted for each patch, and a set of planar parameters for the planar patch is estimated.

In the following sections, we will detail the three components of our approach: patch extraction, stereo matching, and plane merging & parameter refinement

A. Patch and interest point extraction

First, the reference image of the stereo image pair is segmented, using the mean-shift-based approach [1]. 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, its boundary is extracted 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 (with large curvature) between line segments are defined as *interest points*, around which the natural matching primitives are going to be defined in section B.

Because a simple region may only contains a few interest points (for example, a rectangle may only have four interest points on the corners), surface reconstruction may not be accurate if plane fitting only uses four points (since a small error or occlusion can bias the result). Fortunately, vertical lines yield more reliable matches between a pair of images that have horizontal epipolar lines. Therefore, we pick up additional points on the boundary between two consecutive interest points when the line segment connecting them is non-horizontal. As a result, we will have more interest points on vertical lines. Now we are ready to perform the three-step stereo match.

B. Three-step stereo match

Let the left image and the right image be denoted as I1 and I2, respectively. The matching process consists of the following three steps.

Step 1: Global match. In a weak textureless scene, the single pixel based match couldn't work well. So we use a match method based on a group of pixels on an image patch that can be interpreted as a planar patch in 3D. In a typical indoor or 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 image I1, the sum of absolute difference (SAD) is carried out for all pixels on the boundary pair (denoted by IB1 and IB2 for left and right images, respectively) of this region between the two images I1 and I2 with a preset search range. The SAD is defined as

$$SAD(\Delta x) = \sum_{(x,y)\in R} | IB_1(x,y) - IB_2(x+\Delta x, y) | \qquad (1)$$

where the summation is carried out for all the pixels on the boundary of the region (patch) R. The x is the visual displacement (disparity) along the horizontal line, since the stereo images have been rectified. Thus the initial disparity of the region between I1 and I2 is obtained as the one minimizing the SAD. The average minimum match cost (SAD) of each pixel, denoted as Cg, is saved. The closer a plane is to the frontal parallel, the smaller Cg is. The process is very efficient, particularly for a large textureless region, like an indoor wall. The pixels on boundary of patch that lie on the image borders are not taken into account in the global match therefore partially visible regions can also be correctly handled.

Step 2: Local match. Since not all regions are frontal planes in 3D space, the pixels in each region do not have a fixed disparity. Thus, for each interest point, the best match is searched within a neighborhood area of the initial disparity. The neighbor search range R is function of Cg calculated in the previous step

$$R_l = kC_g \qquad (2)$$

where k is a constant. A smaller Cg, means a smaller search range.

Instead of using the conventional window-based match (since it does not work well on the depth boundary if occlusion occurs), 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, called natural window, centered at that interest point such that

$$M(i,j) = \begin{cases} 1, \text{ } if \;\; (x+i, y+j) \in \mathbf{R} \\ 0, \text{ } otherwise \end{cases} \qquad (3)$$

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 (1-3) 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 SAD based on the natural window centered at the point (x, y) in the reference image, is defined as.

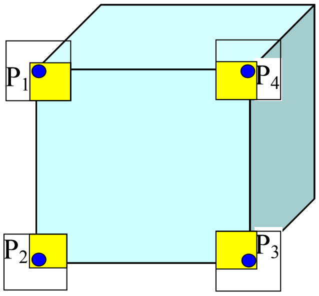

Fig. 3. Natural matching primitives defined in a frontal region.

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

Note that we still carry out sum of absolute difference calculation between two color images, but only on those interest points along each region boundary, and only with those pixels within the region and on the boundaries for each interest point. A confidence value *C* (0<C<=1) is defined as

$$C = \frac{SAD_1}{SAD_2} \qquad (5)$$

where $SAD_1$and SAD_2 are the best (smallest) and the second best match score of each interest point.. The smaller *C* is, the more confidence the match has. 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 a crosscheck followed [7].

Step 3: Surface fitting. Assuming that each homogeneous color region is planar in 3D, a 3D plane

$$aX + bY + cZ + d = 0 \qquad (6)$$

which is represented in the camera coordinate system, is fitted to each region (patch) after obtaining the 3D coordinates of the reliable interest points of the region, since stereo camera head is calibrated (with known intrinsic parameters). We use a robust RANSAC method to fit a plane. RANSAC randomly picks up samples, generates an estimation of the plane and votes from all samples (to either support or not support; each sample votes one ticket). The process is iterative and the estimating result obtaining the most tickets is selected as final estimation. In the voting step, interest points with higher confidences are able to vote more tickets, as

$$T = e^{1/C} \qquad (7)$$

where T is the number of tickets obtained from one interest point, and C is the match cost defined in Eq. (1). Note that at this point the result of a patch after the three steps is in the form of a 3D planar equation and the boundary of each patch.

C. *Plane merging and parameter refinement*

After the above three steps are applied to the pair of stereo images, the estimations of the 3D structures of all the patches (regions) in the reference image are obtained. If the SAD value is less than a preset threshold, then the patch is marked as *reliable*.

Two reasons lead us to merge regions after stereo match. First, we have found that some very 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. Second, one planar surface may be segmented to several sub-regions. In order to recover meaningful surface structures as large as possible for climbing robots, we try to combine them back to one surface. To solve the first problem, we perform a modified version of the neighboring plane parameter hypothesis approach [10] 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. To deal with the second problem, the neighboring regions sharing the same or very close plane parameters are merged into one larger region. This procedure is performed recursively till no more merges occur.

V. Experimental Results

Experiments have been performed to test our approach. Image sequences were captured by the stereovision head Bumblebee, which is fixed on a mobile robot. For a pair of stereo images, the left camera serves as the reference camera. The baseline distance between the left and the right cameras is 12 cm, and focal length of each is 3.8 mm. The stereo system has been pre-calibrated and image pairs rectified.

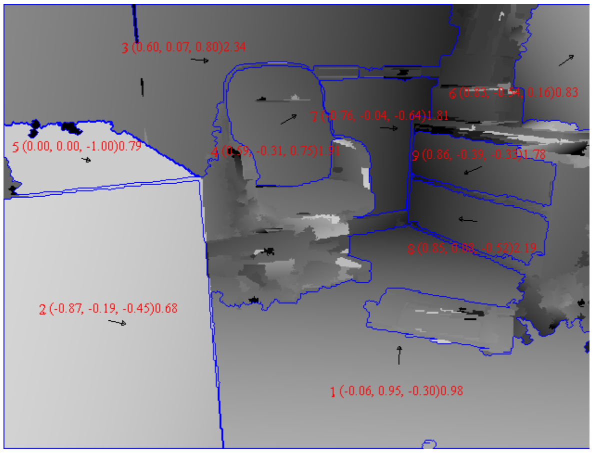

Fig 4. The labeled "depth" map of the first stereo view (the brighter, the closer). For several large regions indexed by 1-9, the boundaries of regions are marked by closed curves (blue) and planar parameters are drawn on the regions: each arrow and the numbers in a pair of parentheses represent the surface norm, and last number (meters) represents the distance from the surface center to the camera.

In Fig. 4, all objects (a table, a chair, a cabinet, walls and some boxes) are about 1-4 meters away from the camera. A pair of color stereo images (Fig. 2.a and Fig. 2.b) and a depth map (the brighter, the closer in Fig. 4) rendered from the result of the plane parametric estimation have been shown. For several large surfaces, the surface norms are visualized by normalized vectors – arrows and values (a,b,c), and distances of the center of the patches to the camera (D in meters), are also labeled, together with their boundaries. These plane estimation results are consistent with the results measured by hand. Readers can visually check the results of plane norms and distances for their correctness. The geometric representations enable a wall-climbing robot to have safe and efficient operations on walls and ceilings. For the textureless regions in the experiment, e.g. the doors, the box and even the ground surface, full 3D results are also obtained.

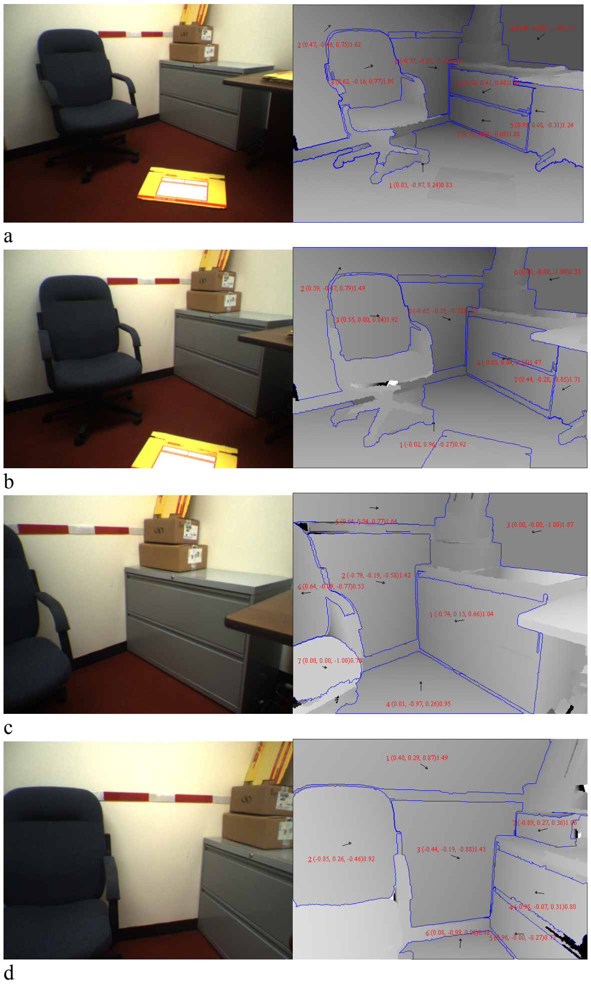

Fig 5. More (closer) stereo views: reference images (left) and corresponding depth maps (right).

In Fig. 5, when the robot moves toward the wall, closer views are captured. From the depth maps, we can see the surfaces are getting closer and closer to the robot.

Surface estimation results have different accuracy based on their distances to the camera. Some of them may not be consistent due to mismatches caused by occlusions and illumination changes. However, in general, the estimation results of a plane are getting more accurate when the camera is closer to the plane. As an example of the stereo accuracy, one pixel match error of a 4-meter away object point will produce a depth error of 0.25 meters, while the same match error only produces a depth error of 0.015 meter if the point is 1 meter away. On the other hand, a larger distance of the scene from the camera (the robot) provides a larger field of view (FOV). Therefore both views are useful for wall-climbing robot navigation and planning. Farther views are good for climbing surface selection, whereas closer views are more accurate for distance and surface orientation estimation.

Fig. 4 and Fig. 5 show results from several stereo views when the robot moves close to the scene. The first view (Fig. 4) can generate a 3D model of planar representation with a relatively large field of view; however, the estimations of distances and orientations of planes of far objects may not be very accurate due to low stereo resolution. For example, the two patches of a file cabinet (No. 8 and No 9 in Fig. 4) have slightly different orientations in the y direction. The closer views (Fig. 5) have relative small space coverage, but with more accurate 3D estimation results. For example, the same regions of the file cabinet are either merged into one planar patch in some views (a to c), or correspond to two patches with very close planar parameters in view (d). In addition, occlusions may also cause errors in the estimation of some planar surfaces. For example, the wall (region 7 in Fig. 4) is partial occluded in different views until the robot moves to Fig 5(d). Even the surface has correct match since the region only select non-occluded boundary to fit the surface (by RANSAC), but it's still not accurate, however, the surface obtains accurate estimate when camera moves to the closer view (Fig. 5d). Theses inconsistencies can be refined by integrating multiple stereo views, which is our ongoing work. The final goal is that the climbing robot not only has global map but also has high accurate plane estimation of planar surfaces of interests to guide the robot to move or climb.

Note that depth information is obtained for all the points in images. Further, the depth information is in the form of parametric representations of the planar surfaces that are ready for wall-climbing robot navigation.

VI. Conclusions and Future Work

In indoor or outdoor urban environments, most of the surfaces of man-made objects are planar, therefore a 3D reconstruction that directly produces plane surfaces is an appropriate approach to solve the problem of visual navigation of a wall-climbing robot working in such environments. In this paper, we have proposed a segmentation-based stereo approach that features natural matching primitives, three-step efficient matching and accuracy parametric 3D estimation. Planar surfaces are represented by their plane parameters (orientations, distances, boundaries), and their relations, which can directly used for visual navigation of our wall climbing robots, the City-Climbers.

After planar representations of the objects at each camera location have been obtained in an unknown 3D environment, the 3D geometric relations of any neighborhood planes in the scene can be built up. Therefore a local topological map of surfaces with 3D metric measures can be created based on the geometric relationships of the planes. When the mobile robot moves, local maps can be integrated into a global 3D map of the unknown environment. This is our ongoing research.

References

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

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

[3] Elliot, M. Morris, W. Xiao, J. "City-Climber, a new generation of wall-climbing robots", *Video Proc. ICRA'06*

[4] Liang, B. and Pears, N. Visual navigation using planar homographies. *ICRA'02*: 205-210.

[5] Lu, X. and Manduchi, R. Detection and Localization of Curbs and Stairways Using Stereo Vision, *ICRA'05*.

[6] Pears, N. and Liang, B. Ground plane segmentation for mobile robot visual navigation, *IROS'01* .

[7] 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

[8] Se, S., Lowe, D., Little, L. Mobile Robot Localization and Mapping with Uncertainty using Scale-Invariant Visual Landmarks, *Int. J. Robotics Research*, 21(8), 2002: 735-758

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

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

[11] Weingarten, J. and Gruener, G. and Siegwart, R. Probabilistic Plane Fitting in 3D and an Application to Robotic Mapping, *ICRA'05*

[12] Xiao, J. Sadegh, A. Elliot, M. A. Calle, A. Persad, H. M. Chiu, Design of Mobile Robots with Wall Climbing Capability, *IEEE/ASME Int. Conf. on Advanced Intelligent Mechatronics*, July 24-28, 2005: 438~443.

[13] Ye, C. and Borenstein, J. A Method for Mobile Robot Navigation on Rough Terrain. *ICRA'0 4:* 3863-3869.