

Robust Scene Reconstruction from an Omnidirectional Vision System

Roland Bunschoten and Ben Kröse

Abstract—In this paper we present an efficient multi-baseline stereo algorithm for panoramic image data. We derive a parameterization of epipolar curves in terms of inverse depth. As a result the search for image correspondences across multiple images can be performed efficiently. Furthermore, depth estimates are obtained directly thus bypassing the need to perform explicit stereoscopic triangulation. We apply our method to obtain a 3D reconstruction of an environment from a set of panoramic images. The images are acquired by a single omni-directional vision sensor mounted on top of our mobile robot during navigation. Experimental results demonstrate the effectiveness of our approach.

Index Terms—Scene reconstruction, multi-baseline stereo vision, omni-directional vision.

I. INTRODUCTION

Recently, researchers in the robotics community have begun to consider omnidirectional vision sensors which provide images covering a large part of the hemisphere. Catadioptric omnidirectional vision sensors are quickly gaining popularity. These sensors consist of a camera and a carefully selected mirror-lens combination. They have been proven to be useful for robot environment modelling, both in the sensory domain (appearance models) [1], [2], [3], [4], as well as in the geometric domain (Cartesian maps) [5], [4].

A traditional approach to obtain a 3D reconstruction of a scene from image data is stereo vision. Stereo vision the process of recovering depth information from two or more calibrated images obtained from different but known camera poses by image based matching and triangulation. Establishing corresponding image points by matching is the fundamental problem in stereo vision (correspondence problem). When the camera poses are unknown (or only approximately known) a priori, they can be estimated from an initial set of corresponding image points. This situation is encountered in our application, where a single camera mounted on top of a moving robot is used to obtain the images.

Particularly useful in stereo applications are catadioptric systems which have single effective viewpoint (see *e.g.* [6], [7], [8] for a concise treatment on such catadioptric systems). Due to the single effective viewpoint property, the epipolar constraint relating two images can be parameterized. The epipolar constraint reduces the correspondence problem from 2D (the entire image domain) to 1D (an epipolar curve in the image domain). Exploiting the epipolar constraint dramatically



Fig. 1

NOMAD SCOUT ROBOT. THE CATADIOPTIC VISION SENSOR IS MOUNTED ON TOP OF THE ROBOT

reduces both the computational demands of image correspondence search as well as the risk of establishing erroneous correspondences. Omnidirectional stereo vision methods using a single pair of images acquired by catadioptric vision sensors have been presented in [9], [10], [11].

In this paper we present an efficient multi-baseline stereo method which reconstructs the 3D environment from a set of images. In our application, omnidirectional images are acquired by a *single* catadioptric vision sensor mounted on top of a mobile robot during navigation. A virtual panoramic vision sensor is constructed by re-projecting an *omnidirectional* image onto a virtual cylinder, yielding a *panoramic* image. Multi-baseline stereo takes advantage of the redundancy contained in the images. A prerequisite for multi-baseline stereo is that the relative camera poses are known. In our application the relative camera poses relating pairs of panoramic images are derived from robot odometry and are refined using tracked image correspondences.

This paper is organized as follows. In section II we describe our catadioptric omnidirectional vision sensor, and the virtual panoramic camera we construct. In section III the epipolar geometry relating two panoramic images is reviewed. Section IV presents a taxonomy of multi-baseline stereo methods. In sec-

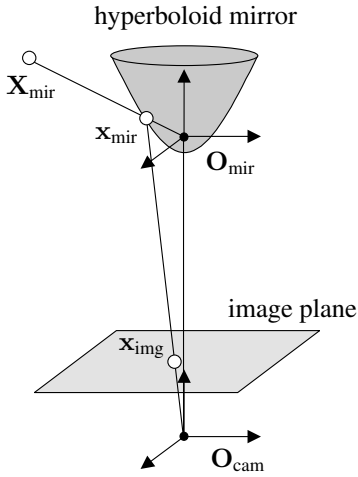


Fig. 2

OMNIDIRECTIONAL IMAGE FORMATION FOR A CATADIOPTIC VISION SENSOR CONSISTING OF A HYPERBOLOID MIRROR AND A PERSPECTIVE CAMERA.

tion V we describe our 3D scene reconstruction method. Experimental results obtained using our method are presented in section VI. A discussion and conclusions are presented in sections VII and VIII.

II. OMNIDIRECTIONAL VISION SYSTEM

Our catadioptric omnidirectional vision sensor consists of a hyperboloid mirror mounted in front of a vertically oriented perspective camera. The hyperboloid mirror has an attractive geometric property illustrated in figure 2: any ray of light which would have passed through the origin O_{mir} of the mirror coordinate frame is reflected such that it passes through the origin O_{cam} of the camera coordinate frame. The position of O_{mir} and O_{cam} are assumed to coincide with the two foci of the hyperboloid. In this way, an omnidirectional vision sensor with a single effective viewpoint is obtained. A detailed description on the design and geometry of the sensor can be found in [12].

The single effective viewpoint property of the omnidirectional sensor enables geometrically correct re-projection of images acquired by the sensor. In our application, we construct a virtual panoramic vision sensor by re-projecting the omnidirectional image onto a virtual cylinder.

Several considerations have led us to use panoramic images instead of omnidirectional images. Whereas a rotation of the robot causes a rotation in the omnidirectional image, it causes a shift in the panoramic image. The correspondence problem can thus be treated without requiring active correlation windows (as proposed in *e.g.* [13]). Off-the-shelf feature trackers, originally developed for conventional perspective cameras, can be employed to obtain an initial set of image correspondences required for camera pose estimation. Finally, the geometry relating panoramic images is simpler than that of the omnidirectional images.

A virtual panoramic camera is constructed by specifying a unit radius virtual cylinder in the mirror frame. The cylinder is

given by $x^2 + y^2 = 1$. Let $\mathbf{X}_{\text{cyl}} = \mathbf{R}\mathbf{X}_{\text{mir}}^T$ be the representation of a point P in the cylinder coordinate frame, where \mathbf{R} is a rotation which aligns the cylinder coordinate frame and the mirror coordinate frame. We define a function $\mathcal{C}(\cdot)$ which computes the central projection of P onto the cylinder surface¹. The projection is computed as the intersection of the ray emitting from the origin of the cylinder coordinate frame and passing through P with the cylinder surface:

$$\mathbf{x} = \mathcal{C}(\mathbf{X}) = \frac{1}{\sqrt{X^2 + Y^2}} \mathbf{X} = \frac{1}{r} \mathbf{X}, \quad (1)$$

where $r = \sqrt{X^2 + Y^2}$. For brevity, we have dropped the subscript indicating that the \mathbf{X} is expressed in the cylinder coordinate frame. Henceforth, we assume that vectors are defined in the cylinder coordinate frame unless explicitly stated otherwise.

We define a function $\mathcal{P}(\cdot)$ and its inverse $\mathcal{P}^{-1}(\cdot)$. Function \mathcal{P} relates the Cartesian coordinate representation $\mathbf{x} = [x, y, z]^T$ of a point on the cylinder surface to its 2D cylindrical coordinate representation $\mathbf{y} = [\phi, z]^T$:

$$\mathbf{x} = \mathcal{P}(\mathbf{y}) = \begin{bmatrix} \cos \phi \\ \sin \phi \\ z \end{bmatrix}. \quad (2)$$

The inverse function \mathcal{P}^{-1} is given by:

$$\mathbf{y} = \mathcal{P}^{-1}(\mathbf{x}) = \mathcal{P}^{-1}(\mathcal{C}(\mathbf{X})) = \begin{bmatrix} \arctan2(y, x) \\ z \end{bmatrix}. \quad (3)$$

The pixels of a virtual panoramic vision sensor are determined by specifying a grid in (ϕ, z) -space. A cylindrical image coordinate \mathbf{y} can be transformed into Cartesian coordinate \mathbf{x} using equation 2. After rotating the Cartesian coordinates into the mirror frame, the image formation steps described in [13] are used to compute the corresponding omnidirectional image coordinate \mathbf{x}_{img} . The intensity value of a panoramic image pixel \mathbf{x} is determined from the intensities in a small neighborhood centered at \mathbf{x}_{img} .

III. EPIPOLAR GEOMETRY FOR PANORAMIC IMAGES

The *epipolar geometry* relates two images obtained by a central projection. The epipolar geometry depends only on the relative pose and internal parameters of the camera(s) by which the images were acquired. Let v_0 and v_1 denote relative camera poses from which two images are acquired by a central projection onto the cylindrical surface. Let $\mathbf{X}_i = [X, Y, Z]^T_i$ denote the coordinate of a scene point P expressed in the i -th camera pose and let \mathbf{x}_i denote its projection onto the imaging surface. We designate v_0 as a reference pose, *i.e.* the coordinate system in which vectors are measured. Let $\mathbf{t}_i = [t_x, t_y, t_z]^T_i$ be the translation vector between the reference pose and the i -th pose and let \mathbf{R}_i be a rotation matrix which aligns the i -th coordinate frame with the reference frame. Then point P can be represented as

$$r_0 \mathbf{x}_0 = \mathbf{t}_i + r_i \mathbf{R}_i \mathbf{x}_i, \quad (4)$$

¹We use uppercase and lowercase characters to distinguish between a point and its projection.

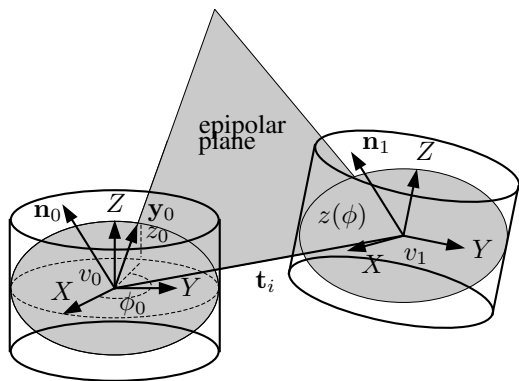


Fig. 3

EPIPOLAR GEOMETRY FOR A PANORAMIC CAMERA.

where r_0 and r_i are unknown depths whose values are to be recovered by stereo.

The epipolar constraint is established by the coplanarity condition

$$\mathbf{x}_i^T \mathbf{R}_i (\mathbf{t}_i \times \mathbf{x}_0) = \mathbf{x}_i^T \mathbf{R} \mathbf{S} \mathbf{x}_0 = \mathbf{x}_i^T \mathbf{E} \mathbf{x}_0 = 0 \quad (5)$$

where \times denotes the vector product, \mathbf{S} denotes the (3×3) skew symmetric matrix such that $\mathbf{S} \mathbf{x}_0 = \mathbf{t}_i \times \mathbf{x}_0$, and $\mathbf{E} = \mathbf{R} \mathbf{S}$ is known as the *essential matrix*.

Figure 3 illustrates that an *epipolar curve* is formed by intersecting the *epipolar plane*, spanned by \mathbf{x}_0 and \mathbf{t}_i , with the cylindrical imaging surface. Given some \mathbf{x}_0 , the search for the corresponding \mathbf{x}_i can be restricted to a search along the epipolar curve. The points where all epipolar curves intersect are called the *epipoles*. They correspond to the direction of motion relating two camera poses. In [9] we have shown that for panoramic images, the epipolar curves are sinusoids.

IV. MULTI-BASELINE STEREO VISION

The correspondence problem is a locally ambiguous problem because distinct points in the scene can have a similar appearance in the images. Errors in the reconstruction obtained from using stereo on a pair of images are therefore likely to be present. Inaccurate reconstruction due to large triangulation uncertainty is obtained near the epipoles. In omnidirectional or panoramic images, these are almost always visible in the image domain.

Multi-baseline stereo vision attempts to overcome these limitations by using more than two images to obtain the reconstruction. A commonly used approach is to first determine corresponding 2D image points across the images. This is then followed by triangulation from image pairs. Combining the depth estimates obtained in this fashion is not trivial; there may be inconsistencies between the depth estimates and re-sampling or interpolation of depth estimates may be required. Several multi-baseline stereo methods have been proposed in literature which address these issues by partitioning the 3D scene space into bins. For each bin a measure of consistency reflecting the likelihood that the bin is occupied is evaluated. We categorize these methods as 2D-3D, 3D-2D or 2D-2D methods.

In 2D-3D methods 3D points are explicitly reconstructed from image correspondences by triangulation. The number of points contained in a bin serves as evidence that the bin is occupied or empty. Triangulation uncertainty can be incorporated so that a reconstructed point does not only contribute evidence to the bin in which it is contained [14].

In 3D-2D methods, 3D bins are projected to multiple images. Seitz and Dyer [15] formulate the scene reconstruction problem as a “voxel coloring” problem. Their method attempts to assign a unique color to each bin that is consistent with all input images. The assumption underlying their approach is that when pixels are back-projected to the same bin, their values should agree. A statistical measure of pixel color consistency is used to determine the occupancy state and color of the bin.

In 2D-2D approaches projective geometry is exploited such that explicit 3D reconstruction of points via triangulation or explicit projection of 3D bins is circumvented. An example is Collins’ “space sweep” approach [16] which exploits the homography between projections of a plane. The approach presented by Okutomi and Kanade [17] exploits a projective invariant called the inverse depth. This scalar quantity expresses depth to a scene point as a fraction of the baseline length and is invariant under changes of the baseline. As a result, depth estimates from different baselines can be combined.

Our approach can be seen as an application of Kanade’s multi-baseline stereo method to panoramic image data and is described in the next section.

V. SCENE RECONSTRUCTION FROM MULTIPLE PANORAMIC IMAGES

In this section we present our scene reconstruction method which reconstructs the environment from a set of K images. In our application the images are acquired by a single omnidirectional camera mounted on top of our mobile robot. Multi-baseline stereo requires that the relative camera poses are known. Section V-A presents the pose estimation method we adopt. In section V-B we present our multi-baseline stereo method for panoramic images.

A. Camera Pose Estimation

A prerequisite of multi-baseline stereo reconstruction is that the camera poses from which images are acquired are known. Although our robot is equipped with fairly accurate odometry, the pose estimates provided by odometry are not accurate enough to fix the epipolar constraint. Due to small errors in the robot orientation estimates, epipolar curves do not pass through corresponding points. Therefore the relative camera poses need to be refined using image correspondences.

An initial set of correspondences is obtained by tracking salient image features through a sequence of images acquired by the robot during navigation. Tracking is performed using KLT [18], a C implementation of the feature tracker described by Shi and Tomasi [19], based on early work of Kanade and Lucas [20]. The tracker fails when the overall image displacement is very large, which occurs when the robot makes a sharp turn. Using odometry measurements, we counter-rotate the virtual panoramic camera such that the overall image displacement

is roughly compensated for and successful tracking can be performed.

After the tracks are obtained, the essential matrix relating each image to the reference image is estimated. For this purpose, we employ a variant of the 8-point algorithm. An iteratively re-weighted least squares estimation procedure [21] (M-estimator) is used to estimate the essential matrix from the remaining set of tracked points. A true essential matrix has two similar eigenvalues and is of rank two. As these conditions are not enforced by the M-estimator, we explicitly enforce them after a solution is obtained as follows. Let σ_1 , σ_2 and σ_3 be the singular values of \mathbf{E} obtained by SVD decomposition $\mathbf{E} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^T$, where \mathbf{U} and \mathbf{V} are rotation matrices and $\mathbf{\Sigma} = \text{diag}(\sigma_1, \sigma_2, \sigma_3)$. We replace $\mathbf{\Sigma}$ by $\hat{\mathbf{\Sigma}} = \text{diag}((\sigma_1 + \sigma_2)/2, (\sigma_1 + \sigma_2)/2, 0)$ and re-compute $\hat{\mathbf{E}} = \mathbf{U}\hat{\mathbf{\Sigma}}\mathbf{V}^T$.

The rotation matrix \mathbf{R} and \mathbf{S} matrix can be determined from the singular value decomposition $\hat{\mathbf{E}} = \hat{\mathbf{U}}\hat{\mathbf{\Sigma}}\hat{\mathbf{V}}^T$ as follows [22]:

$$\mathbf{R} = \hat{\mathbf{U}}\mathbf{Y}\hat{\mathbf{V}}^T \quad \text{or} \quad \hat{\mathbf{U}}\mathbf{Y}^T\hat{\mathbf{V}}^T \quad (6)$$

$$\mathbf{S} = \hat{\mathbf{V}}\mathbf{Z}\hat{\mathbf{V}}^T \quad \text{or} \quad -\hat{\mathbf{V}}\mathbf{Z}\hat{\mathbf{V}}^T \quad (7)$$

where

$$\mathbf{Y} = \begin{bmatrix} 0 & -1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 1 \end{bmatrix} \quad \text{and} \quad \mathbf{Z} = \begin{bmatrix} 0 & -1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}. \quad (8)$$

There are 4 possible combinations of rotation and translation which result in the same essential matrix. The correct combination can be found by recovering the depths to each tracked point according to the relative poses implied by each combination. The correct combination is the one for which most recovered depths are positive.

An estimated essential matrix is only defined up to an arbitrary scale factor. As a consequence, the length of the translation vector relating two images is cannot be determined from image information only. Currently, we use odometry to provide the scale factor.

B. A Multi-Baseline Stereo Method for Panoramic Images

In this section we derive a parameterization of epipolar curves in terms of inverse depth. Using this parameterization, the search for image correspondences across multiple images can be performed efficiently and depth estimates are obtained without explicit stereoscopic triangulation.

Once the camera poses from which images are obtained are known, the virtual panoramic images can easily be re-oriented such that they all have the same orientation. After this rectification, the camera poses are related by translations only. Using equations 2 and 4 the 3D location of a point can be expressed as

$$r_0\mathcal{P}(\mathbf{y}_0) - \mathbf{t}_i = r_i\mathcal{P}(\mathbf{y}_i). \quad (9)$$

Dividing both sides of equation 9 by r_0 and applying the function \mathcal{P}^{-1} (equation 2) which transforms rays to panoramic image coordinates to both sides gives

$$\mathcal{P}^{-1}(\mathcal{P}\mathbf{y}_0 - \frac{1}{r_0}\mathbf{t}_1) = \mathcal{P}^{-1}(\frac{r_i}{r_0}\mathcal{P}\mathbf{y}_i) = \mathbf{y}_i. \quad (10)$$

Note that $\mathcal{P}^{-1}(\cdot)$ eliminates the quantity r_i/r_0 from the right hand side of the equality. Equation 10 shows that the projection of a point on the cylindrical imaging surface at pose v_i is a function of \mathbf{y}_0 , the translation vector \mathbf{t}_i and the fraction $1/r_0$ which is called the inverse depth λ . Using equation 3, the above equation can be expressed in vector form as

$$\begin{bmatrix} \arctan\left(\frac{\sin\phi_0 - \lambda t_y}{\cos\phi_0 - \lambda t_x}\right) \\ \frac{z_0 - \lambda t_z}{\sqrt{(\cos\phi_0 - \lambda t_x)^2 + (\sin\phi_0 - \lambda t_y)^2}} \end{bmatrix} = \begin{bmatrix} \phi_i \\ z_i \end{bmatrix} = \mathbf{y}_i, \quad (11)$$

where $\lambda = 1/r_0$ is the inverse depth quantity.

The above equation is a parameterization of the epipolar curve in terms of inverse depth. It is used to govern the search for image correspondences across multiple images. Given an image coordinate $\mathbf{y}_0 = (\phi_0, z_0)$ and a value for λ , equation 11 gives the image coordinate \mathbf{y}_i corresponding to a scene point at depth $1/\lambda$ from the reference pose.

In our multi-baseline stereo algorithm, for each pixel \mathbf{y}_0 from the reference image, equation 11 is used to generate a set of potentially matching image coordinates \mathbf{y}_i by plugging in multiple values for λ . Subsequently, image similarity is evaluated by computing the sum of squared differences (SSD) between windows centered at \mathbf{y}_0 and \mathbf{y}_i respectively. The SSD values obtained from different images for the same \mathbf{y}_0 and λ are combined by adding them. The underlying assumption is that when an object is present at some depth $r = 1/\lambda$ from the reference pose, the window contents will have a roughly similar appearance in all images, consistently giving rise to small SSD values.

Finally, the most likely depth value λ^* for a specific \mathbf{y}_0 is found by examining the summed SSD values obtained for each λ and selecting the one which yields the smallest value.

Let \mathcal{Y} denote the set of pixel coordinates in the reference image. Let Λ denote a set of inverse depth values whose values are determined to cover a minimal and maximal expected scene depth. The complete multi-baseline stereo algorithm can be summarized as follows:

for all $\mathbf{y}_0 \in \mathcal{Y}$ **do**

$\lambda^* = \lambda_1$

for all $\lambda \in \Lambda$ **do**

$\mathcal{C}(\mathbf{y}_0, \lambda) = 0$

for $i = 1$ to K **do**

compute $\mathbf{y}_i(\mathbf{y}_0, \mathbf{R}_i, \mathbf{t}_i)$ using equation 11

$\mathcal{C}(\mathbf{y}_0, \lambda) \leftarrow \mathcal{C}(\mathbf{y}_0, \lambda) + \text{SSD}(W(\mathbf{y}_0), W(\mathbf{y}_i))$

end for

if $\mathcal{C}(\mathbf{y}_0, \lambda) < \mathcal{C}(\mathbf{y}_0, \lambda^*)$ **then**

$\lambda^* = \lambda$

end if

end for

where $\text{SSD}(W_0(\mathbf{y}_0), W_i(\mathbf{y}_i))$ computes the sum of squared differences between image windows W_0 and W_i centered at \mathbf{y}_0 and \mathbf{y}_i respectively. The map \mathcal{C} contains all summed SSD values. If desired, it is possible to perform some kind of regularization of \mathcal{C} .

VI. EXPERIMENTS

Our experimental platform is a Nomad Scout robot (manufactured by Nomadic Technologies, Inc.) and shown in figure 1.

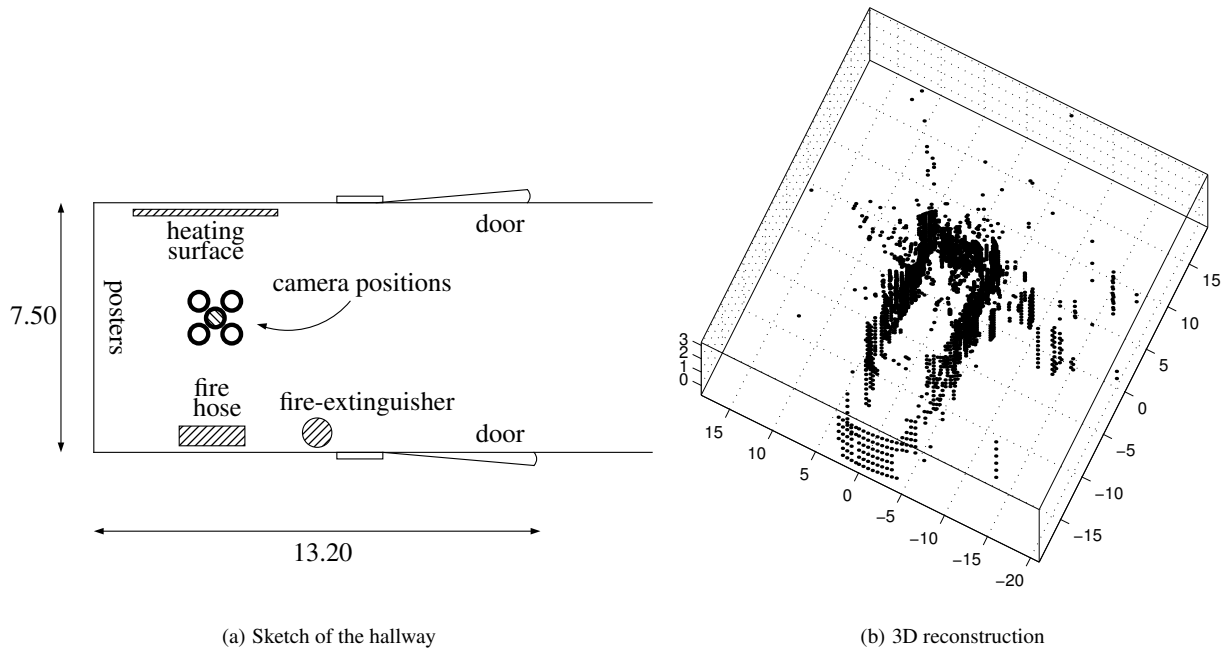


Fig. 4

A) OUTLINE OF THE ENVIRONMENT. FIVE IMAGES WERE ACQUIRED AT THE POSITIONS MARKED BY THE SMALL CIRCLES. THE CENTRAL CIRCLE DENOTES THE LOCATION OF THE REFERENCE IMAGE. B) THE OBTAINED 3D RECONSTRUCTION.



Fig. 5

(TOP) THE REFERENCE IMAGE. (BOTTOM) THE DEPTH MAP COMPUTED FROM THE REFERENCE IMAGE AND 4 OTHER IMAGES.

The robot is equipped with (among other sensors) odometry and an omnidirectional vision sensor. The omnidirectional vision sensor consists of a vertically oriented camera (Sony EVI-370 color camera) and a hyperbolic mirror (manufactured by Accowle, Co., LTD [23]) mounted in front of the camera lens. The hyperboloid omnidirectional images (600×450 pixels) obtained by the vision sensor are transformed into panoramic images (720×120 pixels).

We tested our method at the end of a hallway in our building. A layout of the hallway is shown in figure 4a. This sim-

ple environment lacks large depth discontinuities and occluding objects. Five panoramic images were acquired by the robot. The positions where the images were acquired are indicated by circles displayed in figure 4a. The image obtained at the center was designated as the reference image and is displayed in figure 5. The motion relating the each image to the chosen reference image was estimated using the method presented in section V-A.

A set Λ containing 25 values in the range 0.5–0.05 was used in the stereo search. The set is obtained via an exponential



Fig. 6

(TOP) THE REFERENCE IMAGE. (BOTTOM) A RECTIFIED IMAGE ACQUIRED AT RELATIVE POSE $(-1.21\text{m}, 1.70\text{m}, 102.0^\circ)$ FROM THE REFERENCE POSE.

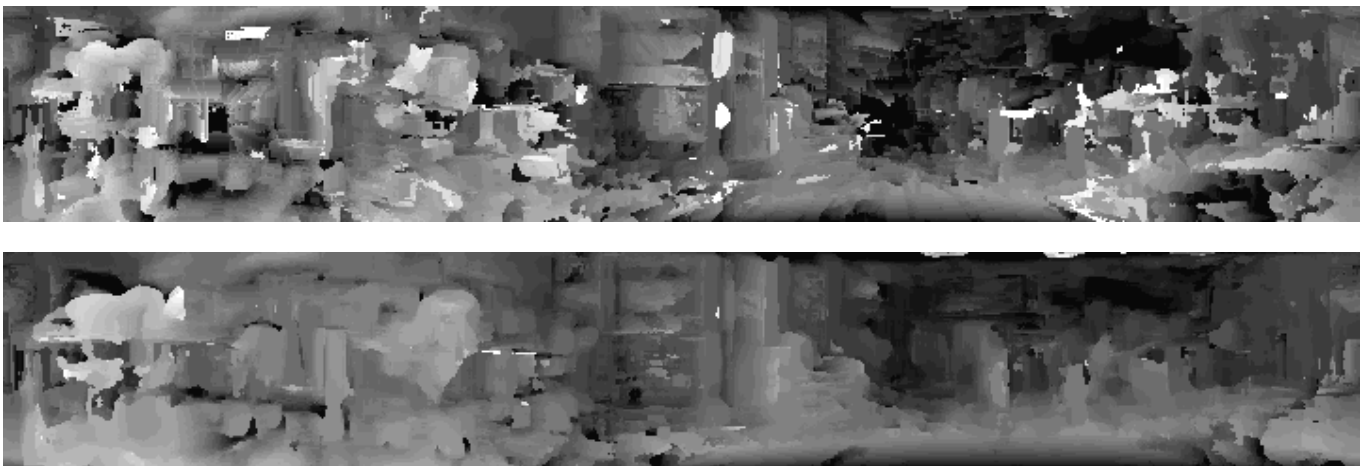


Fig. 7

(TOP) THE DEPTH MAP COMPUTED FROM THE REFERENCE IMAGE AND THE RECTIFIED IMAGE SHOWN IN FIGURE 6. (BOTTOM) THE DEPTH MAP COMPUTED FROM THE REFERENCE IMAGE AND ALL 21 OTHER IMAGES.



Fig. 8

THE REFERENCE IMAGE SHOWN IN FIGURE 6 (TOP) WARPED TO THE IMAGE SHOWN IN FIGURE 6 (BOTTOM).

sampling of depths r in the range 0.5m–20m. By performing exponential sampling, the resulting sampling of points along an epipolar curve is almost uniformly spaced. In the experiment we used 11×11 pixel correlation windows centered at the resulting nearest pixel coordinates to evaluate the sum of squared differences between image windows. Using this correlation window size gave a good tradeoff between accuracy and smoothness of the resulting depth maps.

Figure 5 displays the depth map computed from the reference image and the 4 other images. In the map, nearby objects appear brighter than objects further away from the reference pose. The depth map evidently shows that the general geometry of the hallway is captured well. The far end of the hallway can be recognized near the sides of the depth map. The corners of the hallway at the near end appear slightly darker than the sections of the walls closer to the reference pose. The heating surface is not reconstructed well due to its repeating texture. Furthermore, specular reflections on the posters are not handled well; they appear as large depth discontinuities in the depth map. Due to the lack of texture, only some of the structure of the fire hose is visible in the combined depth map. In figure 4b the 3D reconstruction of the hallway according to the combined depth map is shown. The reconstruction shows that the overall geometry of the hallway is captured well. Gross errors in the reconstruction are caused by lack of texture and repeating texture in the images.

A similar experiment was performed using image acquired in a laboratory with large depth discontinuities and occlusions. A set of 22 images was acquired while the robot traversed a circular trajectory with radius of 1.3m. The first image was designated as the reference image is displayed in figure 6. A rectified image, obtained from the omnidirectional image captured at relative pose $(-1.21m, 1.70m, 102.0^\circ)$, is shown in figure 6. The depth map computed from the reference image and the other image is shown in figure 7. As can be observed, the depth map contains many errors (arising due to occlusions, lack of texture, repeating texture, specular reflections etc). Such noisy depth maps are typically obtained from panoramic image pairs. Application of our multi-baseline stereo technique improves the estimated depth maps. In figure 7 the depth map computed from the reference image and all other image is shown. Overall, the geometry of the environment is captured well but some gross errors remain. These errors occur mainly at locations where there are large depth discontinuities and occlusions (such as the chair on the left side in the images) and where there is little texture (such as on the floor).

Equation 11 can also be used to warp the reference image to a target image that would be obtained at a target position. In figure 8 we show the reconstruction of the image displayed in figure 6 obtained by warping the reference image according to depth map obtained from all 21 image pairs and the known target camera pose. A forward mapping scheme was used to warp the reference image. The forward mapping leaves holes in the warped image. In order to render a visually more appealing image, the holes were filled by interpolating gray values from neighboring pixels.

VII. DISCUSSION

The depth maps obtained by our multi-baseline stereo method are good, especially when considering that no regularization of depth maps is performed. In our current implementation, a fixed set of 25 λ 's is used to compute depth maps. We are currently working on an extension of our method which maintains a small set of λ 's for each pixel from the reference image independently. The idea is to maintain only those inverse depth values which are likely to correspond to an object in the scene point (based on previously estimated map). Using gradient information, refinements of inverse depth can be obtained by the method of differences [20]. This approach resembles conditional density propagation by particle filtering, which has been successfully applied in the field of visual object tracking [24] and mobile robot localization [25], [26].

Another direction we investigate is the use of equation 11 in the context of appearance based environment modelling. A drawback of appearance based environment representations is that many training images are needed to obtain an accurate model. An approach to overcome this problem was presented in [27], where based on a number of measured range profiles synthetic profiles are generated. In a similar manner, image warping can be used to generate synthetic *images* from measured images.

VIII. CONCLUSION

We have presented a scene reconstruction algorithm for panoramic images obtained by a single moving panoramic vision sensor. We have derived a parameterization of the epipolar curve in terms of inverse depth. Using this parameterization the search for 2D image correspondences and the 3D reconstruction from multiple images can be performed efficiently. A depth map obtained from a single image pair is very noisy. The improvement that can be achieved using multi-baseline stereo has been demonstrated by experiments.

REFERENCES

- [1] B. Kröse, R. Bunschoten, N. Vlassis, and Y. Motomura, "Appearance based robot localization," in *IJCAI-99 Workshop on Adaptive Spatial Representations of Dynamic Environments*, G. Kraetzschmar, Ed., 1999, pp. 53–58.
- [2] B. Kröse, N. Vlassis, R. Bunschoten, and Y. Motomura, "A probabilistic model for appearance-based robot localization," *Image and Vision Computing*, vol. 19, no. 6, pp. 381–391, Apr. 2001.
- [3] M. Jogan and A. Leonardis, "Panoramic eigenimages for spatial localisation," in *Proc. of the 8th Int. Conf. on Computer Analysis of Images and Patterns*. 1999, number 1689 in LNCS, pp. 558–567, Springer Verlag.
- [4] N. Winters and J. Santos-Victor, "Mobile robot navigation using omnidirectional vision," in *Proc. 3rd Irish Machine Vision and Image Processing Conf.*, Dublin, Ireland, Sept. 1999.
- [5] Y. Yagi, K. Shouya, and M. Yachida, "Environmental map generation and egomotion estimation in a dynamic environment for an omnidirectional image sensor," in *Proc. IEEE Int. Conf. on Robotics and Automation*, San Francisco, 2000, pp. 3493–3498.
- [6] T. Svoboda, T. Pajdla, and V. Hlaváč, "Central panoramic cameras: Geometry and design," Tech. Rep. K335/97/147, Center for Machine Perception, Czech Technical Univ., 1997.
- [7] S.K. Nayar and S. Baker, "A theory of catadioptric image formation," Tech. Rep. CUCS-015-097, Dept. of Computer Science, Columbia Univ., 1997.
- [8] C. Geyer and K. Daniilidis, "A unifying theory for central panoramic systems and practical applications," in *Proc. European Conf. on Computer Vision*, Dublin, Ireland, 2000.

- [9] R. Bunschoten and B. Kröse, "Range estimation from a pair of omnidirectional images," in *Proc. IEEE Int. Conf. on Robotics and Automation*, Seoul, Korea, May 2001, pp. 1174–1179.
- [10] J. Gluckman, S.K. Nayar, and K.J. Thoresz, "Real-time omnidirectional and panoramic stereo," in *Proc. Image Understanding Workshop*, 1998.
- [11] H. Ishiguro, M. Yamamoto, and S. Tsuji, "Omni-directional stereo," *Pattern Analysis and Machine Intelligence*, vol. 14, no. 2, pp. 257–262, February 1992.
- [12] T. Pajdla, T. Svoboda, and V. Hlavac, "Epipolar geometry of central panoramic cameras," in *Panoramic Vision : Sensors, Theory, and Applications*, Ryad Benosman and Sing Bing Kang, Eds., pp. 85–114. Springer Verlag, Berlin, Germany, 1st edition, 2001.
- [13] T. Svoboda and T. Pajdla, "Matching in catadioptric images with appropriate windows and outliers removal," in *Proc. 9th Int. Conf. Computer Analysis of Images and Patterns*, Berlin, Germany, September 2001, pp. 733–740, Springer Verlag.
- [14] H.P. Moravec, "Robot spatial perception by stereoscopic vision and 3-D evidence grids," Tech. Rep. CMU-RI-TR-96-34, The Robotics Institute, Carnegie Mellon Univ., Pittsburgh, Pennsylvania, Sept. 1996.
- [15] S.M. Seitz and C.R. Dyer, "Photorealistic scene reconstruction by voxel coloring," in *Proc. Int. Conf. Computer Vision and Pattern Recognition*, 1997, pp. 1067–1073.
- [16] R.T. Collins, "A space-sweep approach to true multi-image matching," in *Proc. Conf. Computer Vision and Pattern Recognition*, 1996, pp. 358–363.
- [17] M. Okutomi and T. Kanade, "A multiple-baseline stereo," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 15, no. 4, pp. 353–363, April 1993.
- [18] S. Birchfield, "KLT: An implementation of the kanade-lucas-tomasi feature tracker," available at: <http://vision.stanford.edu/~birch/klt/>.
- [19] J. Shi and C. Tomasi, "Good features to track," in *IEEE Conference on Computer Vision and Pattern Recognition*, June 1994, pp. 593–600.
- [20] B.D. Lucas and T. Kanade, "An iterative image registration technique with an application to stereo vision," in *Proc. Imaging Understanding Workshop*, 1981, pp. 121–130.
- [21] P.H.S. Torr and D.W. Murray, "The development and comparison of robust methods for estimating the fundamental matrix," *Int. J. of Computer Vision*, vol. 24, no. 3, pp. 271–300, 1997.
- [22] R.I. Hartley, "Estimation of relative camera positions from uncalibrated cameras," in *2nd European Conf. on Computer Vision*. May 1992, LNCS 588, pp. 579–587, Springer-Verlag.
- [23] H. Ishiguro, "Development of low-cost and compact omnidirectional vision sensors," in *Panoramic Vision: Sensors, Theory and Applications*, R. Benosman and S.B. Kang, Eds., pp. 35–41. Springer Verlag, 2001.
- [24] A. Blake and M. Isard, "The CONDENSATION algorithm — conditional density propagation and applications to visual tracking," in *Advances in Neural Information Processing Systems*, Michael C. Mozer, Michael I. Jordan, and Thomas Petsche, Eds. 1997, vol. 9, p. 361, The MIT Press.
- [25] D. Fox, W. Burgard, F. Dellaert, and S. Thrun, "Monte carlo localization: Efficient position estimation for mobile robots," in *Proc. 6th National Conf. on Artificial Intelligence; Proc. 11th Conf. on Innovative Applications of Artificial Intelligence*, Menlo Park, Cal., July 18–22 1999, pp. 343–349, AAAI/MIT Press.
- [26] N. Vlassis, B. Terwijn, and B. Kröse, "Auxiliary particle filter robot localization from high-dimensional sensor observations," in *Proc. IEEE Int. Conf. on Robotics and Automation*, Washington, D.C., May 2002, To appear.
- [27] J.L. Crowley, F. Wallner, and B. Sciele, "Position estimation using principal components of range data," in *Proc. IEEE Int. Conf. on Robotics and Automation*, 1998, pp. 3121–3128.