Proceedings of the 2002 IEEE
International Conference on Robotics & Automation
Washington, DC • May 2002

Toward Practical Cooperative Stereo for Robotic Colonies.

Bart Nabbe and Martial Hebert
The Robotics Institute
Carnegie Mellon University
Pittsburgh, PA, 15217, USA
{bana,hebert}@cs.cmu.edu

Abstract— **In this paper we describe an approach towards cooperative stereo[1]. The key problems, significant different views, different scale and occlusion, will be addressed in the context of a distributed robotic system.**

An algorithmic approach is described towards solving these problems and results on real scenes are presented.

Keywords— **Wide Baseline Stereo; Affine Invariants; Robust Epipolar Estimation**

I. Introduction

Recovery of scene structure at long distances, *e.g.*, hundreds of meters is critical for mobile robot navigation in expansive outdoor environments. Structure recovery at such distances is not possible from fixed, on-board stereo systems because a much wider baseline is needed than can be achieved on a single robot. Generally speaking, very wide baseline between camera views can be achieved either by accumulating images over long distances, *i.e.*, structure from motion (SFM), or by using images taken from widely separated robots, *i.e.*, cooperative stereo. SFM is often favored because it simplifies the matching problem by using small incremental motion between images. In many cases, however, it is not possible to accumulate image data over a sufficiently long baseline to recover structure accurately. This is particular true for structure recovery in the direction of motion. Furthermore, in cluttered environments, it is difficult to track a large enough set of features to recover 3-D structure.

To address the potential shortcomings of SFM in robotics systems, we investigate in this paper the feasibility of the alternative approach, cooperative stereo, to perform image matching over the very wide baselines necessary for long-range structure recovery. This approach is attractive because it makes no assumption on the robot's motion and, in principle, requires only images from a few positions of the robots. Recent advances in wide-baseline matching in the Computer Vision literature provide a solid basis to tackle this problem. In particular, the landmark work of Zisserman and his colleagues [1] provides the basic tools for wide-baseline stereo. Our work is based in large part on that formulation; we extend it and evaluate it in the context of very wide baselines for robotic mapping.

For a mobile robot control standpoint, recent results in multi-robot cooperation [2] show that multi-robot planners can support the coordination of multiple robots to ensure coverage of a scene by their sensors. It is important to note that we focus here on the design and performance of cooperative stereo algorithms understanding that, in a complete mobile robot system, both cooperative stereo and SFM should be used. Integrating both approaches to take advantage of their respective strengths remains to be done.

A. Cooperative stereo

The motivation for this work is the development of a distributed robotic system which purpose is to map unknown environments. This distributed system consists of several robots, each equipped with a limited set of sensors. Each robot has only a single camera for mapping purposes.

A complete mapping task would be subcontracted out and bid on by individual robots, using the *Free Market Architecture* [2] and [3]. In order to make a stereo measurement, a single robot would have to cooperate with another robot to take the second image. In a setup like this, there is only a limited amount of control over the camera position. It is therefore necessary for the stereo algorithm to deal with significant different views and occlusions, different scale and camera rotation. In this context, we want to perform

a b c

Fig. 1. A few typical outdoor scenes.

[1]This work was supported by the Army Research Labs Robotics Collaborative Technology Alliance Contract DAAD 19-01-2-0012

cooperative stereo in three broad classes of environments, as shown in Figure 1. First, we want to address environments that contain a lot of structure, such as man-made environments, as shown in Figure 1(a), at moderate distances, *e.g.*, up to 100m. This type of environment is the closest to the examples typically studied in the existing work on wide-baseline stereo. The second class of environments is shown in Figure 1(b) in which little regular structure can be extracted from the images. In such cases, we need to rely on ill-defined features such as regions uniform in color or texture rather than on geometric features such as straight lines and corners. A third case is shown in Figure 1(c) in which the scene does contain a fair amount of structure but the distance to the camera and the baseline between the robots is considerably larger. The large aspect difference between the images induced by the very wide baseline complicates stereo matching substantially. The environments of Figure 1(b) and Figure 1(c) depart from the type of environments normally used in wide-baseline stereo work and are closer to target environments for colonies of mobile robots operating in outdoor environments.

B. Wide baseline stereo techniques

The key to cooperative stereo is the recovery of the epipolar geometry of two widely separated cameras. Traditionally, the epipolar geometry is estimated by matching local features. More precisely, if I_1 and I_2 are two images separated by rotation and translation $\boldsymbol{R}, \boldsymbol{t}$, the relative poses of the cameras are recovered by minimizing with respect to $\boldsymbol{R}$ and $\boldsymbol{t}$ an error of the form $\sum_i d(F_i^1, T_{\boldsymbol{R},\boldsymbol{t}}(F_i^2))$, where F_i^1 and F_i^2 are corresponding features in images I_1 and I_2, and $T_{\boldsymbol{R},\boldsymbol{t}}(F_i^2))$ is the projection of F_i^2 in I_1 given pose $\boldsymbol{R}, \boldsymbol{t}$. For example, if the features are point features, d is the usual distance between image points and epipolar lines. If the features are planar patches, $T_{\boldsymbol{R},\boldsymbol{t}}$ is a homography and d is the correlation between image patches.

We focus in this discussion on the recovery of $(\boldsymbol{R}, \boldsymbol{t})$ from the images. Once the epipolar geometry is known, a variety of techniques can be used for recovering the structure. The 3-D positions of individual features can be recovered by direct triangulation [4]. A number of techniques can be used for recovering dense structure, *e.g.*, [5], [6]. Therefore, we concentrate here on the crucial step of recovery of the epipolar geometry.

The challenge in wide-baseline stereo is to choose the features so that they remain sufficiently invariant over a wide enough range of camera poses to be matched. These features must be descriptive enough such that they can be distinguished from each other and at the same time reproducible and invariant to the several image transfer functions. Furthermore, because the percentage of acceptable matches decreases rapidly as the viewpoints become more separated, the design of robust matching and numerical estimation algorithms become critical.

A first possible type of feature is a corner detector and an associated local image descriptor. A comparison between the feature descriptor of both images will yield candidate matches. A robust estimator is then used to find a subset of matches that support a consistent epipolar geometry. This is the approach that was used in the early work on the recovery of epipolar geometry [7], in which direct correlation of templates centered at the interest points is used to establish initial correspondences between images.

For wide-baseline matching, the variation in the appearance of the the neighborhoods of the interest points is so great that direct correlation cannot be used for matching. Instead, more involved local image descriptors that remain invariant under a wide range of image transformations are used. Schmid and Mohr [8] developed such a descriptor that is rotational invariant and can be made scale-invariant by using a multi-resolution representation. Local image descriptors that are invariant to affine transformations can be constructed based on the approach of [9] in which local templates are warped to a canonical, *i.e.*, invariant, normalized frame. Templates centered at features from multiple images are all represented in the normalized frame and can be matched in an affine-invariant fashion. This is the approach taken in [10].

Techniques based solely on interest points tend to be limited to scenes with a moderate degree of occlusion. The reason is that a point feature is typically found on the occluding boundaries of objects. The variation of local appearance of such a point between two widely separated images is due not only to geometric warping between the two images, which can be corrected by proper use of the invariance properties, but also by variation in the occlusion geometry due to the change of viewpoint. The second source of appearance variation dominates in practice and cannot be compensated by geometric invariance (Figure 2.)

Fig. 2. Local appearance of an interest point detected at an occluding boundary, viewed from two widely separated viewpoints. The primary source of appearance variation is due to the change in occlusion geometry and cannot be compensated by warping the window.

One way to circumvent this weakness of point features is to use extended regions instead of individual point features. This is the approach used in [11] in

which affine-invariant moments are computed from regions extracted from the images. Initial matches are established between regions with similar moments. A RANSAC approach [7] is used to filter out the initial matches and to estimate the final epipolar geometry. In this case, regions are extracted by first finding local extrema in the images and by defining a region at a fixed radius around each extremum. This effectively avoids extracting features near occlusions.

The use of invariant descriptors is effective at compensating for the variation of appearance of features and regions between images. There is a price to pay, however: the features become less discriminative because information is lost in the process of making the features invariant. In addition, features that impose strong constraints on $(\boldsymbol{R}, \boldsymbol{t})$ are preferable, so that fewer matches are needed to recover camera geometry. In particular, matching two planar patches in two images defines a homography between the images and two such matches are, in principle, sufficient for recovering camera geometry [12]. The use of planar patches as features for wide-baseline stereo was proposed in [13] in which planes are constructed from sets of corners and the homographies computed from corresponding planes are used to find additional point matches. Spurious matches are filtered out using a RANSAC technique as before.

This approach still relies on the detection of point features. In unstructured environments, however, point detection is unreliable. A more promising approach would be to extract salient regions from a segmentation of the images, and to use those regions for matches. This is the approach proposed in [1], in which the normalized cuts segmentation approach [14] is used to generate the initial regions. Each region of the segmentation is warped to a normalized frame, and seed points in these regions are used for matching and recovery of the epipolar geometry.

This work is closest to our approach to cooperative stereo. We also use a segmentation algorithm to generate initial regions and planar homography to constrain matches, but our approach differs in two key ways: First, to take advantage of the presence of strong geometric structure in many environments, we also use groupings of lines in the image to extract regions that can be used as possible planar patches for matching whenever possible. Second, we use direct correlation between regions in order to refine the transformations - homographies - between regions. This approach allows us to take advantage of all the texture information contained in a region. In contrast, previous approaches use point correspondences to refine the transformations.

Another difference is that, we combine filtering out of spurious matches and estimation of camera geometry into a single numerical criterion as instead of RANSAC. Although a RANSAC-like method could be used as well. Finally, we evaluated the algorithms on scenes (Figure 1) that are challenging in terms of viewpoint variation, scene complexity, or lack of geometric features. Experimentation of cooperative stereo techniques on such scenes is critical for future practical applications to mobile robotics.

II. Algorithm description

In this section, we describe the algorithm for recovering epipolar geometry and scene structure, from widely separated viewpoints. After extracting features from the two images, as described in the next section, we apply progressively tighter filtering criteria to the set of candidate matches generated from the two sets of features from the two images.

The overall strategy for finding a set of matching regions can be summarized as follows. Each region i from image j, R_i^j, is represented by a feature vector f_i^j describing global characteristics of the regions, *e.g.*, color histogram. f_i^j is used for establishing initial matches between regions based on global properties. In addition, each R_i^j is warped to a new image R_i^{jo} in a normalized frame, *e.g.* a fixed square patch, using a homography $\boldsymbol{H}^{oj}_i$ computed from the region's parameter. If two image regions R_k^1 and R_l^2 correspond to the same, approximately planar, physical patch, then R^{oj}_k and R^{oj}_l should be similar. This fact is used to further refine the set of matches between regions. Finally, given a candidate set of matches, the homography H_n mapping the first region $R^1{}_{k_n}$ of correspondence n to the second region $R^2{}_{l_n}$ is estimated, using the homography induced by $\boldsymbol{H}^{o1}_{k_n}$ and $\boldsymbol{H}^{o2}_{l_n}$. The approach and the notations are summarized in Figure 3.

Below, we describe the three levels of filtering and refinement of the set of candidate matches, leading to the final evaluation of the epipolar geometry.

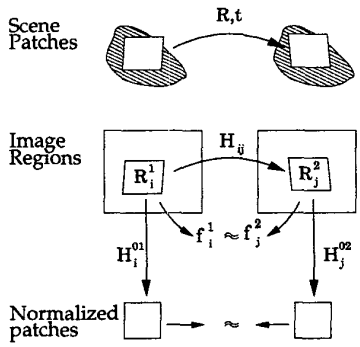

Fig. 3. General approach to finding matching regions.

A. Feature Extraction

The scenes we consider can be roughly divided in two classes: mostly geometric man-made and natural scenes. The approaches taken for these scenes differ only in the first phase of the algorithm. In both cases, we have ruled out approaches based solely on point features in favor of region-based techniques for the reasons stated above. Similarly, we rely on the approximate planarity of the image regions to reduce the number of matches needed for geometry recovery and to further constrain matching.

Feature Extraction: Line Groupings

It has been shown, in the context of 3-D modeling from multiple aerial images [15], that, for scenes with substantial geometric structure content, regions can be extracted by grouping line segments. Four line segments are normally needed for each region, although, in practice, three line segments are sufficient with the fourth one being inferred. To avoid searching through all possible 3-tuples of line segments, the endpoints of the line segments are first stored in bins equally spaced in the horizontal and vertical directions. Bins of 50-pixel width are used. For each segment, candidate segments with connecting endpoints are retrieved by consulting the appropriate bins. Line segments are divided into two categories: near-horizontal and near-vertical lines. The two groups are defined based on large tolerance, $\pm$ 20°, between the lines and the horizontal and vertical directions. Groups are formed by finding connecting lines segments from the two categories. The use of the image axis as reference to aid in grouping restricts somewhat the generality of the approach. In practice, for images taken by mobile robots in which in-plane camera rotation is limited, the number of groups that are missed is small enough that it has no effect on the performance of the system. The result of this initial grouping approach is a set of cycles which is further filtered based on thresholds on the ratio of lengths of opposing line segments and on the elongation of the regions enclosed by the cycles.

Each region R_i^j corresponding to a cycle is mapped to a normalized reference patch by computing the transformation H^{oj}_i that maps its vertices to a 15x15 square. The intensity values (Y channel) in the region defined by the cycle are mapped to the normalized patch. Such a small size is used for the normalized patch so that comparison of regions for initial matching is fast. It is important to note that the normalized patches are used only for filtering out correspondences; the transformations between regions are estimated by using all the pixels within the regions.

Figure 4 shows a typical set of line groupings and an example of normalized patch representation. Typical for these real images, a number of spurious cycles are detected and the normalized patches may not match even if the groups are geometrically correct. Filtering and refinement of the matches are described in the sections below.

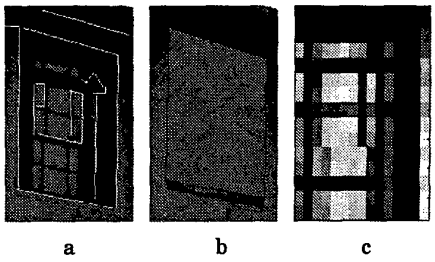

Fig. 4. Lines extracted from an image (a), a hypothesized line grouping (b) and the reference normalized patch (c)

Each region R_i^j is represented by a feature vector f_i^j which includes:

- Area and axes of the region. The axes are computed by taking the eigenvectors of the second moment matrix of the region.
- A color histogram computed in the U and V channels of the YUV color-space. Each channel was converted into a 3bit color-depth. These two vectors were concatenated in a 64-bin histogram.

This feature vector is used for generating initial matches as described later in this section.

Feature Extraction: Regions

A color segmenter [16] is used in natural scenes that lack strong geometric features. The same feature vectors as before, area and axes, and color histogram, are used for representing each region. Using the same feature vectors as in the case of regions computed from line groupings allows us to use exactly the same machinery for matching.

In addition, each region is warped to a normalized patch as in the case of the four-cycle representation above. More precisely, each segmented region is mapped to a reference square patch. The transformation is estimated by mapping the two axis and the two elongations of the second moment matrix to the reference axes.

Figure 5 shows a typical example of a segment from an image segmentation. The ellipse computed from the second-order matrices is shown as is the normalized reference frame. As can be seen in this example, many regions are segmented incorrectly due to oversegmentation, occlusion, and photometric variations. This is typical of real outdoor images. It is the job of the matching algorithm described below to retain those few regions that can be matched reliably. This underscores the importance of using homographies to reduce the number of features that need to be correctly matched in order to estimate camera geometry. Also, the mapping from region to normalized patch is

only an approximation since the boundary of a region varies substantially due to instabilities in segmentation. This motivates the need for the refinement step described below to compute the actual homography between regions.

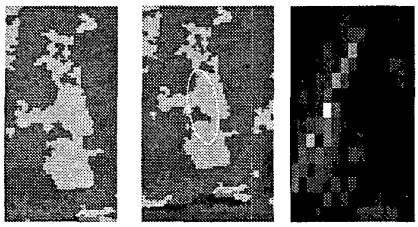

Fig. 5. a typical segment from an image segmentation (a), the principal axis (b) and the normalized reference patch (c).

B. Initial Filtering of Matches

Using the feature descriptor computed as described above, a first filtering step is applied to eliminate incorrect matches. Due to the large number of possible matches, this first step uses low-computation tests. Also, liberal thresholds are used to avoid filtering out acceptable matches at this stage.

We first reject all matches that have a significant mismatch in area, axis length and orientation. Matches are rejected if the area is off by a factor of 10, an axis mismatch by a factor of 3 or a 30° orientation difference. In order to further reduce the candidate matches, we only keep the 20 best color matches. These are the matches with the smallest sum of squared differences (L_2-distance) of the two color histograms.

This initial filtering reduces the number of potential matches typically to a dozen or so out of an initial set on order of one thousand candidates.

C. Initial Matching

For the remaining candidates the normalized gray-scale patch is used to compute a similarity measure. For this, we will use the result, as presented by Pritchett and Zisserman [13], that cross-correlation is geometrically invariant if the pixels of two planar patches are mapped unto each other under a homography. So if the result of a cross-correlation between these warped patches yield a high correlation score, the underlying assumption that the patches are the same and planar holds.

To compute an initial assignment, we would like to find the best global consistent pairing between features in the left and right image. To do this, we extend the approach suggested in [17], [18], [19] in the context of matching point features to planar regions. In this approach, we define a cross-correlation similarity measure as a cost metric weighted by a proximity term [17]. This proximity term G_{ij} will favor an assignment with a global minimal displacement.

$$G_{ij} = e^{\frac{-r_{ij}^2}{2\sigma^2}} \tag{1}$$

With r_{ij} the Euclidean distance between the feature in the first and the second image. This distance metric follows a Gaussian decay over the support region σ. This distribution favors more distant matches, but because of the gradual decay and the large support region, matches within a shorter range are still accounted for. For our experiments we have found that a σ of $\frac{1}{7}$th of the image width is sufficiently large to capture the typical disparity range. The new total cost for a match between feature i from one image and j from the other is given by:

$$G_{ij} = 1 - \frac{(C_{ij}+1)}{2} e^{\frac{-r_{ij}^2}{2\sigma^2}} \tag{2}$$

In which C_{ij} is the normalized cross-correlation value between patch i and j. The costs in this cost matrix G ranges from 0 to 1, the smaller the cost, the better the match. In order to be able to assign a one to one mapping between the two feature sets, the cost matrix is padded to a square matrix with the rogue feature match cost set to a constant value (0.4).

The matching problem now follows the regime of a maximum weighted matching problem for bipartite graphs for which a standard solution is the *Hungarian method* [20], a polynomial-time algorithm. The output of this stage is a set of possible correspondences between regions, which has been filtered based on similarity between feature vectors, similarity between normalized patches, and global consistency.

Figure 6 shows the set of candidate matches retained at this point in the matching algorithm for two regions.

D. Refinement

Since the number of potential correspondences is reduced to a small number N by using the steps above, it is now possible to use more expensive algorithms to compute the exact transformation between regions. In keeping with our general approach, we use all the pixels in the regions to estimate the transformation rather than estimating the transformation from sets of feature points extracted inside the regions. This may be potentially more expensive than using point matches, but it leads to more accurate estimates of the transformations. This is in contrast with other wide-baseline stereo work [1] in which the transformations are estimated from point features.

To simplify notations, we denote by (R_i^1, R_i^2), $i = 1, \ldots, N$, the set of matches identified at the previous

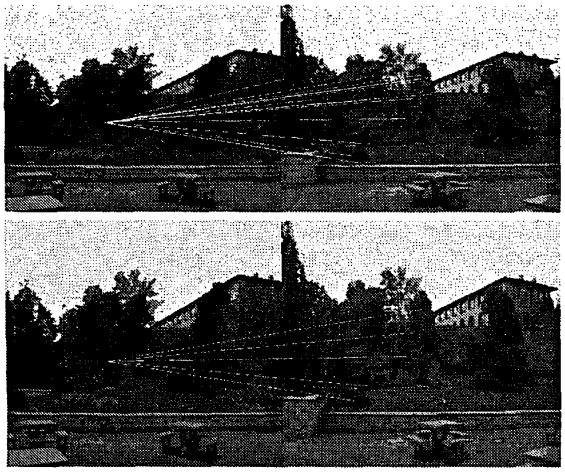

Fig. 6. Unfiltered matches (top), candidate matches after filtering based on feature vectors and normalized patch similarity (bottom).

steps. Since those candidate matches have fairly high confidence, we can now compute explicitly the transformation $\boldsymbol{H}_i$ that maps R_i^1 to R_i^2 for every correspondence i. $\boldsymbol{H}_i$ is recovered by minimizing the objective function J defined as:

$$J(\boldsymbol{H}_i) = \sum_{\boldsymbol{x} \in R_i^1} (I_2(\boldsymbol{H}_i \boldsymbol{x}) - I_1(\boldsymbol{x}))^2, \qquad (3)$$

where $\boldsymbol{H}_i$ the homography that maps region R_i^1 in the first image onto region R_i^2 in the second image. Levenberg-Marquardt is used to adjust the homography $\boldsymbol{H}_i$ such that the similarity between the patch in the second image and the warped patch from the first image is optimal. The LM iterations are initialized at the value of $\boldsymbol{H}_i$ induced by the homographies to the normalized patches, $\boldsymbol{H}^{o1}_i$ and $\boldsymbol{H}^{o2}_i$.

In practice, this type of optimization is known to be able to recover only small pixel discrepancies between patches (strictly speaking, at most one pixel motion can be recovered.) In practice, however, the initial homography computed from $\boldsymbol{H}^{o1}_i$ and $\boldsymbol{H}^{o2}_i$ may be relatively far from the optimum due, for example, to instability in region segmentation. To allow for a larger discrepancy between the regions, we can minimize the J first using lower resolution of the images and use the resulting $\boldsymbol{H}_i$ to start the estimation at the next higher level of resolution. Four levels of resolution seem sufficient to handle the typical errors between matching regions. The images are blurred between resolution levels by a Gaussian of $\sigma = 11$.

E. Initial epipolar estimation

The homographies $\boldsymbol{H}_i$, $N = 1, \ldots, N$, estimated as described above are used for computing the initial epipolar geometry. As before, while the homographies can be used to find point correspondences [21], we prefer to first use the homographies directly to compute a first estimate without using additional point features. If $(\boldsymbol{R}, \boldsymbol{t})$ is the transformation between the two viewpoints, the epipolar geometry is characterized by $\boldsymbol{E} = [\boldsymbol{t}]_\times \boldsymbol{R}$. This is assuming that the cameras are calibrated and that the image coordinates are expressed as normalized coordinates. Classical results from multi-view geometry show that the epipolar geometry can be recovered exactly from two homographies [12].

Based on this classical result, we could pick pairs of homographies from the set $\boldsymbol{H}_i$, $i = 1, \ldots, N$ and compute the corresponding epipoles and epipolar geometry. A better approach is to use all the homographies simultaneously to compute $\boldsymbol{t}$ and $\boldsymbol{R}$. We use the standard representation of each homography as:

$\boldsymbol{H}_i = \lambda_i \boldsymbol{R} + \boldsymbol{t}\boldsymbol{v}_i^T$,

where $\boldsymbol{v}$ is the scaled normal to plane i: $\boldsymbol{v}_i = \lambda_i \frac{\boldsymbol{n}_i}{d_i}$ (the translation is normalized to unit length by convention.) We then find the unknowns, $\boldsymbol{R}, \boldsymbol{t}, \lambda_i, \boldsymbol{v}_i, i = 1, \ldots, N$ that minimizes the difference between $\boldsymbol{H}_i$ and $\lambda_i \boldsymbol{R} + \boldsymbol{t}\boldsymbol{v}_i^T$, summed over $i = 1, \ldots, N$. It turns out that the solution to this problem can be computed by first considering a virtual homography: $\boldsymbol{H} = \sum_i \boldsymbol{H}_i = \lambda \boldsymbol{R} + \boldsymbol{t}\boldsymbol{v}^T$, where $\boldsymbol{v}$ is the scaled normal to the virtual plane. If $\boldsymbol{H} = \boldsymbol{U}\boldsymbol{S}\boldsymbol{V}^T$ is the SVD decomposition of $\boldsymbol{H}$, then the optimal $\boldsymbol{R}$ and $\boldsymbol{t}$ is obtained by equating λ to the largest singular value of $\boldsymbol{H}$ and $\boldsymbol{t} = \boldsymbol{U}\boldsymbol{t}', \boldsymbol{v} = \boldsymbol{V}\boldsymbol{v}'$. The components of the vectors $\boldsymbol{t}'$ and $\boldsymbol{v}'$ are computed as solutions of fourth-order polynomial equations whose parameters depend only of $\boldsymbol{S}$. This approach is similar to the one used in [22].

F. Robust estimation

The previous step provides an estimate of $\boldsymbol{R}$ and $\boldsymbol{t}$ from the homographies. This estimation was performed so far without outlier rejection so that the current estimate may be corrupted by spurious matches. As a last step, to eliminate possible remaining spurious matches, we re-estimate $\boldsymbol{R}$ and $\boldsymbol{t}$ based on the vertices of the regions. We denote by $\boldsymbol{P}^1_{ij}$, $j = 1, \ldots, 4$ are the four vertices of region i in image 1 and $\boldsymbol{P}^2_{ij} = \boldsymbol{H}_i \boldsymbol{P}^1_{ij}$ are the vertices in image 2 obtained by transformation by the current estimate of the homography $\boldsymbol{H}_i$. In the case of region extraction from lines, the vertices are the vertices of the polygon defining the region; in the case of free-form segmentation, the vertices are extremities of the principal axes of the second-moment matrix of

the region. $\boldsymbol{R}$ and $\boldsymbol{t}$ are estimated by minimizing:

$$E(\boldsymbol{R},\boldsymbol{t}) = \sum_{i=1,j=1}^{i=N,j=4} \psi(d(\boldsymbol{P}^2_{ij}, \boldsymbol{E}\boldsymbol{P}^1_{ij}) + d(\boldsymbol{P}^1_{ij}, \boldsymbol{E}^T\boldsymbol{P}^2_{ij})), \tag{4}$$

where $d()$ is the usual normalized distance between image point and epipolar line, and $\psi()$ is a Lorentzian function of the form $\psi(d) = \frac{d^2}{\sigma^2+d^2}$. The scale σ is proportional to the median of the error $d()$ over all pairs of points. This approach is the one used, for example, in robust registration [23]. An alternative robust estimation technique would be to use a technique such as RANSAC. In practice, the outliers are effectively eliminated by using a Lorentzian directly, which also has the advantage to produce an optimal estimate of $(\boldsymbol{R},\boldsymbol{t})$ at the same time.

III. Experimental results

For the typical images as show in Figure 1, we have included the epipolar reconstruction results.

In Figure 7 the result for structured scene is shown. The result for this type of scene is very good as can be seen by the precise epipolar reconstruction. This is further illustrated in Figure 10 by the fact that the ratio between the edges of two reconstructed windows matches fairly closely: 1.5456 for the left and 1.8356 for the right. The algorithm initially considered all the 262x305 matches, after filtering only 369 of these matches were considered for refinement and 41 matches passed the global matching and robust optimization phase.

The result for the more natural scene (Figure 8) was computed using only the patches from the segmentation stage. These 533x919 original patches were filtered down to 137 candidate matches, from which 17 made it through the global matching and robust optimization phase. Because the segmentations are much influenced by noise, this result is far less accurate as the previous result.

A much more exciting result is shown in Figure 9 here we show a result that was computed from a 33m baseline. Locations in the scene cover a distance from a 100m to a 1000m. A small error has therefore quite some influence in the result. Nevertheless the reconstruction is fairly accurate. Here we started out with considering 982x669 matches, after filtering we had 1107 matches left from which 194 survived the global matching and robust optimization.

IV. Conclusion

We show in this paper that a more geometric feature descriptor in conjunction with an extensive filtering pipeline can be used to solve the cooperative stereo problem. The results as presented, are steps towards a fully cooperative stereo approach for robotic colonies. Further work will involve actively positioning robots to create an adjustable baseline and integration of midrange stereo data into the on-board planner.

Fig. 10. The ratio of the height over the width of the window on the left is 1.5456 and the window on the right 1.8356, which gives us an indication of the reconstruction quality.

References

[1] F. Schaffalitzky and A. Zisserman, "Viewpoint invariant texture matching and wide baseline stereo," in *Proc. 8th International Conference on Computer Vision, Vancouver, Canada*, July 2001.

[2] M. Dias and A. Stentz, "A free market architecture for distributed control of a multirobot system," in *Proceedings of the 6 th Intl. Conf. on Intelligent Autonomous Systems, Venice, Italy*, July 2000.

[3] S. Thayer, B. Digney, M. Dias, A. Stentz, B. Nabbe, and M. Hebert, "Distributed robotic mapping of extreme environments," in *Proceedings of SPIE: Mobile Robots XV and Telemanipulator and Telepresence Technologies VII*, November 2000.

[4] Richard I. Hartley and Peter Sturm, "Triangulation," *Computer Vision and Image Understanding*, vol. 68, no. 2, pp. 146–157, 1997.

[5] O. Veksler, *Efficient Graph-based Energy Minimization Methods in Computer Vision*, Ph.D. thesis, Cornell University, July 1999.

[6] Stan Birchfield and Carlo Tomasi, "Multiway cut for stereo and motion with slanted surfaces," in *Proc. International Conference on Computer Vision, Kerkyra, Greece*, September 1999.

[7] Zhengyou Zhang, Rachid Deriche, Olivier Faugeras, and Quang-Tuan Luong, "A robust technique for matching two uncalibrated images through the recovery of the unknown epipolar geometry,," *Artificial Intelligence, December 1995*, vol. 78, pp. 87–119, 1995.

[8] C. Schmid and R. Mohr, "Local gray-value invariants for image retrieval," in *IEEE Transactions on Pattern Analysis and Machine Intelligence, 19(5)*, May 1997.

[9] T. Lindeberg and J. Gøarding, "Shape-adapted smoothing in estimation of 3-d shape cues from affine deformations of local 2-d brightness structure," *Image and Vision Computing*, vol. 15(6), pp. 415–434, June 1997.

[10] A. Baumberg, "Reliable feature matching across widely separated views," in *IEEE Conf. on Computer Vision and Pattern Recognition*, 2000.

[11] T. Tuytelaars and L. Van Gool, "Wide baseline stereo matching based on local affinely invariant regions," in *Proc. British Machine Vision Conference*, 2000.

[12] R. Hartley and A. Zisserman, *Multiple View Geometry in Computer Vision*, Cambridge University Press, 2000., 2000.

[13] P. Pritchett and A. Zisserman, "Wide baseline stereo matching," in *Proc. 6th International Conference on Computer Vision, Bombay*, 1998.

Fig. 7. The reconstructed epipolar geometry from the structured environment.

Fig. 8. The reconstructed epipolar geometry from the more natural environment.

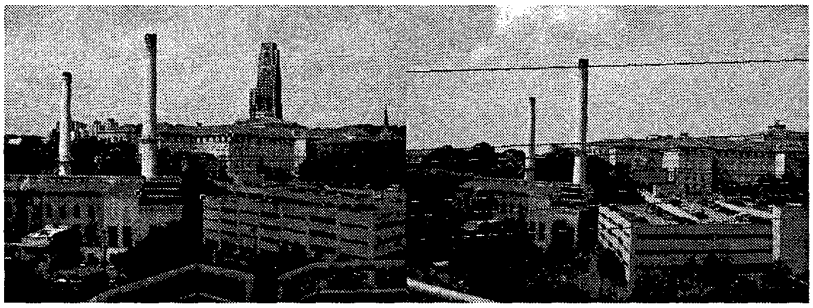

Fig. 9. The reconstructed epipolar geometry from the very wide baseline

[14] J. Shi and J. Malik, "Normalized cuts and image segmentation," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 22, no. 8, pp. 888–905, 2000.

[15] C. Baillard and A. Zisserman, "Automatic reconstruction of piecewise planar models from multiple views," in *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, 1999.

[16] D. Comaniciu and P. Meer, "Robust analysis of feature spaces: Color image segmentation," in *IEEE Conf. on Computer Vision and Pattern Recognition, Puerto Rico*, 1997.

[17] M. Pilu, "A direct method for stereo correspondence based on singular value decomposition," in *IEEE Conf. on Computer Vision and Pattern Recognition*, 1997.

[18] G. Scott and H. Longuet-Higgins, "An algorithm for associating the features of two images," *Royal*, vol. B-244, pp. 21–26, 1991.

[19] L. Shapiro and J. Brady, "Feature-based correspondence: An eigenvector approach," *IVC*, vol. 10, no. 5, pp. 283–288, June 1992.

[20] G. Fielding and M. Kam, "Applying the hungarian method to stereo matching," in *Proc. 1997 IEEE Conference on Decision and Control*, December 1997.

[21] A. Bartoli, P. Sturm, and R. Horaud, "Projective structure and motion from two views of a piecewise planar scene," in *Proc. IEEE International Conference on Computer Vision*, 2001.

[22] O. Faugeras and F. Lustman, "Motion and structure from motion in a piecewise planar enviroment," *Journal of Pattern Recognition and AI*, vol. 2(3), pp. 485–508, 1988.

[23] M.D. Wheeler and K. Ikeuchi, "Sensor modeling, probabilistic hypotheses generation, and robust localization," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 17, no. 3, 1995.

[24] J. Malik, S. Belongie, J. Shi, and T. Leung, "contours and regions: Cue combination in image segmentation," in *Proc. International Conference on Computer Vision, Kerkyra, Greece*, September 1999.