

FUSING VIDEO AND SPARSE DEPTH DATA IN STRUCTURE FROM MOTION

Qilong Zhang and Robert Pless

Department of Computer Science and Engineering
Washington University in St. Louis
St. Louis, MO. 63130
{zql, pless}@cse.wustl.edu

ABSTRACT

This paper considers the geometric constraints to combine structure from motion with a sparse set of depth measurements. The goal is to improve the motion estimation for autonomous navigation, and to increase the fidelity of reconstructed 3D scene models. The system is implemented on an iRobot-B21r Robot with a video camera and a planar laser range finder which gives relatively accurate depth measurements of a small set of scene points. Using a probabilistic model of scene smoothness, the depth information is used to modify the classical epipolar error function to simultaneously incorporate data from both sensors. We present the results of real-world experiments and experiment with different prior assumptions about the scene structure.

1. INTRODUCTION

This paper presents an integrated algorithm for calculating the camera motion and scene structure from a system with a camera and a planar laser range finder. The structure from motion (SFM) problem (sometimes called the structure *and* motion problem to emphasize that both are initially unknown) is a classical problem in computer vision. Recently, a large body of literature shows that the SFM problem is inherently unstable, and in real world situations there is a large amount of uncertainty in both the camera motion and the reconstructed scene [1, 2].

Although an enormous amount of work has gone into making purely visual algorithms robust, the robotics community, with its focus on practical and reliable systems, has turned to systems that integrate many sensors. In this vein, systems that measure scene properties use other sensors, commonly laser scanners (e.g. [3]), or laser range finders. These systems capture one or more range images and then merge them into a complete 3D scene representation (e.g. [4]). Image information is then used to define the texture or reflectance properties for the 3D model. These methods of 3D reconstruction are expensive — either in terms of the time taken for the laser to scan across the scene or in terms

of the cost of purchasing or creating a system that can scan multiple image areas at once.

However, it is becoming more common to have limited range sensing capabilities, and this leads us to consider the problem addressed in our paper: Define a structure and motion algorithm that uses both image and sparse range information. Our implementation and experiments use the planar laser range finder that comes with many standard research robots. However, the specifics of the range sensor are *not* important, the algorithm will work with any sensor or system that gives depth information about (a few of the) points in the scene. This could include recognizing a known object in the scene (and therefore determining depth by relating the known object size and the image size), a stereo system that returns the distances to a few points in the scene with definite correspondences, or depth computed from just a few passes of a laser scanner.

The specific geometric problem considered here is: *Given corresponding points in two images and a sparse set of image points with known depth, compute the rigid transformation that relates the two images and reconstruct the scene structure.* The two goals of this paper are:

- Develop an integrated algorithm that simultaneously uses the sparse depth measurements and the image correspondences.
- Measure the effectiveness of sparse depth information in improving the SFM computation.

Our algorithm can be expressed simply in terms of a modification to the standard epipolar constraint equation used in vision algorithms. Our algorithm is presented explicitly in Section 3. Intuitively, we define a range estimate and a relative confidence for every image pixel. The confidence is based upon the correlation between the depth of nearby pixels which has been demonstrated experimentally [5]. This correlation decreases as the distance between image pixels increases, so image pixels that are far from any range estimate will have a low confidence. In these regions, the SFM computation is effectively image based.

The following section presents some mathematical background and notation, which allows the algorithm to be described. Section 4 gives examples of experimental results. We also present a modified version of the algorithm that exploits regularities in the range image common to indoor scenes.

2. BACKGROUND

A 2D point is denoted by $m = [u, v]^T$, and a 3D point is denoted by $P = [x, y, z]^T$. We use \tilde{x} to denote the homogeneous representation of a vector x . For a usual pinhole camera, a projection from a world point P to the image point m can be represented as following [6]:

$$\tilde{m} \sim K[R \ t]\tilde{P} \quad (1)$$

where K is the camera intrinsic matrix, R a 3×3 orthonormal matrix representing the camera's orientation, and t a 3-vector representing its position. In real cases, the camera can exhibit significant lens distortion. We assume in the remainder of this paper that the camera has no significant lens distortion, or that the images have already been warped to eliminate it.

Considering the case of a calibrated camera moving in a static scene, images are acquired in two consecutive time instants, which are respectively denoted by I and I' . Let (R, t) represent the camera displacement between these two time instants, and m and m' be the images of a 3D point P . The two image points satisfy the epipolar constraint

$$\tilde{m}'^T F \tilde{m} = 0 \quad (2)$$

where F is the fundamental matrix relating image I and I' . If the camera motion (R, t) is known, we can directly calculate fundamental matrix F as following:

$$F = K^{-T}[t]_{\times} R K^{-1} \quad (3)$$

where $[t]_{\times}$ is the skew-symmetric matrix defined by t .

Suppose that a set of image correspondences $m_i \leftrightarrow m'_i$ are given from image I and I' . And these image correspondences are the projections of a set of 3D points P_i . Without loss of generality, we assume that a 3D point P is expressed in the camera coordinate frame at the first time instant. Assuming a calibrated camera (which implies that K from Equation 3 is known), the common approach to SFM solves for the fundamental matrix F by minimizing the error function:

$$\sum (\tilde{m}_i^T F \tilde{m}_i)^2, \quad (4)$$

then factoring F to get estimates for the rotation R and the translation t , and using those estimates as initial conditions in a global optimization to find the $(R, t, P_1, \dots, P_i, \dots)$ which minimize:

$$\sum (\|m_i - \varphi(P_i)\|^2 + \|m'_i - \varphi(RP_i + t)\|^2) \quad (5)$$

where $\varphi(\cdot)$ is the camera projection function corresponding to Equation 1.

When the intrinsic camera calibration K is unknown, the scene structure may be reconstructed up to an unknown projective distortion. Small errors in the camera calibration lead to uncertainty in the reconstruction, and the even without errors in the calibration function the convergence of minimization techniques for these error functions is slow. Thus, for robust algorithms, it is important to fuse other sensor data, such as range data, with the constraints from vision systems.

The range sensor reports range data which are distance measurements to points in the scene, measured from the coordinate system of the range sensor. This coordinate system is related to the camera coordinate system by some rigid transformation. Solving for this calibration is heterogeneous sensor calibration problem and there are calibration algorithms based on calibration pattern [7]. In the remainder of this paper, we assume that the camera is registered to the range sensor coordinate system, and the 3D points from the laser range finder have been transformed to the camera coordinate system.

The proposed method proceeds by exploiting range information acquired simultaneously with camera image capture, as robot moves in space. Provided with additional depth information in regions overlapping with visual data, the epipolar equations are augmented with depth constraints, which improve the robustness of SFM computation.

3. ALGORITHM DESCRIPTION

This section provides the details how to fuse range with visual data in SFM problem. The algorithm gives an integrated approach to this fusion problem, and there is no requirement that the sparse set of image points where the depth is measured $\{p_j\}$ overlaps with the set of corresponding points $\{m_i, m'_i\}$.

For a normalized¹ image point m_i of image I , with a known depth Z_i , a hypothetical 3D point can be expressed as $\tilde{m}_i Z_i$. We first present one method to estimate Z_i , then how to integrate that estimate, and its confidence, with the epipolar error measure.

Given a set of image points $\{p_j\}$ with known depth, we choose the point p_j closest on the image plane to m_i . The closest point minimizes the squared Euclidean image distance $\|p_j - m_i\|^2$. A study of the statistics of range images [5] states that nearby pixels are more likely to have similar depths — the confidence that the depth of the scene at point m_i is similar to the depth of the point p_j (one of the

¹Normalized images clarify the geometric presentation. A normalized image assumes the camera calibration matrix K is the identity matrix. For any known camera calibration the equivalent normalized image can be created

sparse set of points where the depth is actually measured) is: $e^{-(p_j - m_i)^\top \Lambda^{-1} (p_j - m_i)}$, for a covariance matrix which we choose to be a multiple of the identity matrix. The magnitude of the diagonal elements of Λ depends on the variability of the scene. In our experiments we have empirically chosen values, a more extensive statistical analysis would give a more reasoned approach to setting this value.

This depth estimate allows us to use a reprojection error, specifically (similar to Equation 1):

$$\sum (\|m'_i - \varphi(R \tilde{m}_i Z_i + t)\|^2) \quad (6)$$

where $\tilde{m}_i Z_i$ is the hypothetical 3D point corresponding to using the depth estimate nearest m_i . This error function forms the addition component to the epipolar error. It is combined with the classical epipolar error with a weight function that depends on the confidence in the depth estimate as described in the integrated epipolar error function defined below.

The reprojection error has a concrete interpretation as the image distance between the reprojected point of $\tilde{m}_i Z_i$ and the actual corresponding point m'_i found in the second image. The epipolar error given in Equation 4 is an algebraic distance. To integrate these errors, it is important to redefine the epipolar error as: $d^2(m'_i, F\tilde{m}_i)$, the squared Euclidean distance between the image point and the its corresponding epipolar line [6]. With these modifications, and making the error function symmetric so that neither image is dominant, leads to the *integrated epipolar and sparse range error function*:

$$\begin{aligned} & \sum (d^2(m'_i, F\tilde{m}_i) + d^2(m_i, F^\top \tilde{m}'_i)) + \\ & \sum (\lambda_i \|\varphi(R \tilde{m}_i Z_i + t) - m'_i\|^2 + \\ & \lambda'_i \|\varphi(R^{-1}(\tilde{m}'_i Z'_i - t)) - m_i\|^2) \end{aligned} \quad (7)$$

What is left is to define λ , a scalar weight determining the relative importance of the laser data and the image data. If the image positions of corresponding points are measured with Gaussian error, then the squared epipolar distance function $d^2(.,.)$ is equivalent to the negative-log-likelihood that the points are accurate correspondence. Similarly, the probability that nearby points have the same depth is Gaussian, then $(p_j - m_i)^\top \Lambda^{-1} (p_j - m_i)$ is the negative log-likelihood that the image point and the nearest sparse range point have the same depth. Choosing $\lambda \sim \frac{1}{(p_j - m_i)^\top \Lambda^{-1} (p_j - m_i)}$ gives a weighting function that is large when the image point corresponds to the same parts of the scene and is small when there is no nearby range point (and, therefore, the confidence in the depth estimate is small).

The point of the above optimization on camera ego-motion relies on using as much depth information provided by range sensor as possible while meeting the epipolar constraint on visual data from camera. In our experiments we solve this

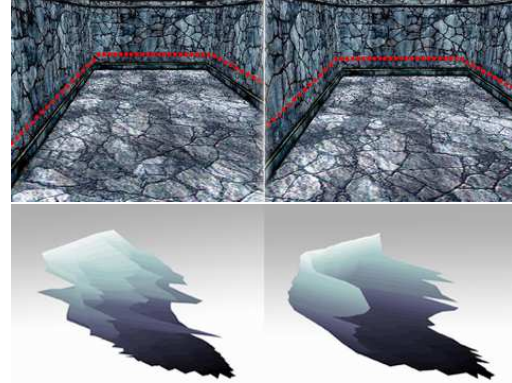


Fig. 1. (top) Two consecutive frames of a synthetic scene with overlaying red points presenting the projections of 2D range measurements acquired by the range sensor. (bottom) The left is the reconstruction using just image correspondences, and on the right using the integrated algorithm, fusing the image and the range data

nonlinear minimization problem with the Levenberg-Marquardt algorithm. The next section illustrates simulated and real-world results, and shows that the use of very sparse depth data significantly accelerates the convergence of these nonlinear optimization.

Summary of the Algorithm: We now summarize the main steps of our proposed algorithm:

1. Acquire images and range data simultaneously in two consecutive time instants, as robot moves in a static scene.
2. Extract feature points from images and compute the point correspondences $m_i \leftrightarrow m'_i$. This was implemented with a stereo correspondence algorithm based on singular value decomposition [8] with RANSAC.
3. Estimate camera motion (R, t) by minimizing the integrated epipolar constraint Equation 7
4. Reconstruct the scene structure by global optimization of Equation 5. Use computed 3D coordinates of corresponding image points and the range finder points to define depth map.

4. EXPERIMENTS

The proposed method has been tested on both synthetic and real data. Figure 1 shows two consecutive frames of computer synthetic scene, together with overlaying red points representing 2D range measurements reported by a laser range finder.

First we present SFM results using corresponding points alone, compared with the integrated image and depth constraints using range data. Figure 1(bottom) shows on the

left the reconstruction using just visual data, and on the right by fusing range information. A large number of post-processing techniques exist to improve the results of SFM from image correspondences, but to emphasize the improvements from using sparse laser measurements we present unaltered reconstructions that are directly the result of minimizing Equation 5.

In order to quantify the improvement in convergence, we also ran simulated studies on how the number of sparse range data is related to the improvement. Figure 2 shows the convergence time for the optimization function which shows the advantage of even a few range points.

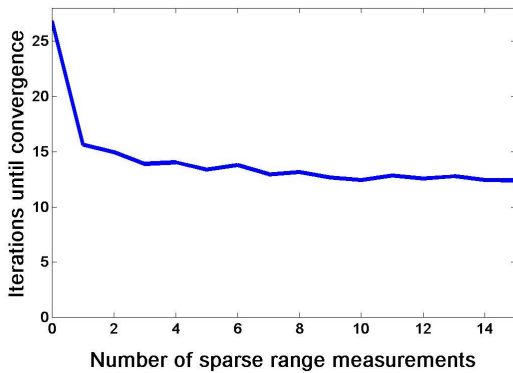


Fig. 2. A plot of convergence time in iterations as more sparse range estimates are available. Note that almost all improvement come with very few points. This shows that sparse range measurements are valuable for SFM estimation.

The proposed algorithm has been implemented and performed on a robotic platform iRobot-B21r, equipped with a SICK-PLS laser range finder and a SONY DFW-VL500 digital camera at a resolution of 640×480 . The range finder has a viewing angle of 180° and reports planar range scans of the surroundings at the height of 18 inches paralleled to the ground plane, with an angle resolution of one measurement per degree and a range measuring accuracy of 5cm. A calibration procedure has been applied, and the relative position of both sensors were known, as well as their intrinsic parameters.

Figure 3 demonstrates the algorithm applied on the corner scene inside our lab. Using only image correspondences, small errors in the motion estimation lead to significant errors in the scene reconstruction, both in the small scale (the roughness of the reconstructed surface), and in the large scale (the incorrect global scene geometry). The scene reconstruction computed with the integrated epipolar constraint solves both these problems.

Conclusion: This paper presents an integrated algorithms for fusing range and visual data for structure from motion. Experimental results indicate the very sparse depth estimates (at just a few points in the image) substantially increase the convergence speed for the optimization problem. Further

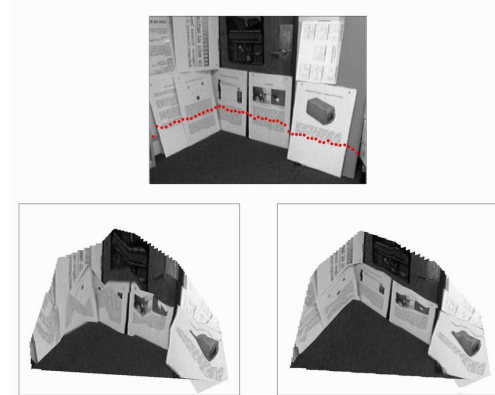


Fig. 3. The image on top shows one frame of lab corner scene captured by the camera, with projections of planar laser range data. The bottom two images show on the left the reconstruction using just image data, and on the right fusing 2D range data

work on understanding the statistics of natural range images will allow more optimal fusion rules.

5. REFERENCES

- [1] K. Daniilidis and H. Nagel, “Analytical results on error sensitivity of motion estimation from two views,” *Image and Vision Computing*, vol. 8, pp. 297–303, 1990.
- [2] J. Oliensis, “A new structure from motion ambiguity,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 22, no. 7, pp. 685–700, 2000.
- [3] M. Levoy, K. Pulli, and et al., “The digital michelangelo project: 3D scanning of large statues,” in *SIGGRAPH, Computer Graphics Proceedings*, 2000, pp. 131–144.
- [4] A. Hilton, A. Stoddart, J. Illingworth, and T. Windeatt, “Reliable surface reconstruction from multiple range images,” in *Proc. European Conference on Computer Vision*, 1996, pp. 117–126.
- [5] J. Huang, A. Lee, and D. Mumford, “Statistics of range images,” in *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, 2000, pp. 324–331.
- [6] R. Hartley and A. Zisserman., *Multiple view geometry in computer vision*, Cambridge University Press, 2000.
- [7] Q. Zhang and R. Pless, “Extrinsic auto-calibration of a camera and laser range finder,” Tech. Rep., Washington University, 2003.
- [8] M. Pilu, “A direct method for stereo correspondence based on singular value decomposition,” in *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, 1997, pp. 261–266.