Long-term motion estimation from images

Dennis Strelow[1] and Sanjiv Singh[2]

[1] Google, Mountain View, CA, strelow@google.com
[2] Carnegie Mellon University, Pittsburgh, PA, ssingh@cmu.edu

Summary. Cameras are promising sensors for estimating the motion of autonomous vehicles without GPS and for automatic scene modeling. Furthermore, a wide variety of shape-from-motion algorithms exist for simultaneously estimating the camera's six degree of freedom motion and the three-dimension structure of the scene, without prior assumptions about the camera's motion or an existing map of the scene.

However, existing shape-from-motion algorithms do not address the problem of accumulated long-term drift in the estimated motion and scene structure, which is critical in autonomous vehicle applications. The paper introduces a proof of concept system that exploits a new tracker, the variable state dimension filter (VSDF), and SIFT keypoints to recognize previously visited locations and limit drift in long-term camera motion estimates. The performance of this system on an extended image sequence is described.

1 Introduction

1.1 Overview

Cameras are small, light, and inexpensive; do not project energy into the environment; are not inherently limited in range; and can be used to estimate motion without an absolute positioning device like GPS. For these reasons, they are good sensor candidates for estimating the motion of micro air vehicles, Mars rovers, and search and rescue robots; and for automatic scene modeling.

Furthermore, a variety of shape-from-motion algorithms exist for estimating camera motion and sparse scene structure from the camera's image stream, including bundle adjustment, the factorization method, the iterated extended Kalman filter (IEKF), and the variable state dimension filter (VSDF). But at the same time, no hands-free systems exist for estimating long-term vehicle motion or modeling complex environments from images. In existing systems, errors in feature tracking and camera intrinsics calibration, degenerate camera motions such as pure rotation, and poor assumptions on the camera motion

can lead to gross local errors in the estimated motion. In addition, existing systems do not explicitly consider the problem of limiting drift during long-term motion.

This paper describes a system for long-term shape-from-motion, i.e., for simultaneously estimating long-term camera motion and scene structure from images, without any assumptions about the camera's motion or an existing map of the camera's environment. This system recognizes when a location visible earlier in the image sequence has been revisited, and limits drift in the current motion estimate by "rolling back" to and extending a previous motion estimate for that location.

1.2 Related work

Simultaneous localization and mapping (SLAM) algorithms typically estimate three degree of freedom (planar) sensor motion and x, y scene structure without prior knowledge of either, from range measurements and odometry. These methods attempt to limit drift in long-term estimates by recognizing previously visited locations and creating topologically correct estimates. Montemerlo, *et al.*[2] describe a recent example.

On the other hand, there has been little work on extending shape-from-motion to long sequences by recognizing previously visited locations or creating topologically correct estimates. In addition, shape-from-motion algorithms recover six degree of freedom camera motion and three-dimensional scene structure from projections (i.e., two-dimensional image feature measurements), without odometry. So, they recover more unknowns from less generous measurements. In addition, the resulting estimates can be highly sensitive to errors in the projections and camera calibration, which often results in gross errors in the estimated motion even for short ("local") sequences. Strelow discusses this issue in more detail ([6], section 5.3.4).

But, recent vision work addresses these problems in part. Nister, *et al.*[3] exploit a fast five-point shape-from-motion algorithm and RANSAC to generate robust and accurate local motion and structure estimates in real-time. By concatenating these estimates, the authors are able to generate accurate motion and structure estimates even for some long sequences. This system is able to generate three-dimensional motion and structure from image sequences alone. But, the system is ad hoc in that its estimates are not optimal in any sense. In addition, this system does not recognize previously visited locations, so the estimates will necessarily drift over time.

Lowe's SIFT keypoints and feature vectors[1] reliably extract and match salient points over wide changes in camera viewpoint, and are a promising approach for comparing widely spaced images in a sequence. Se, *et al.*[4], describe a system that uses SIFT keypoints, stereo cameras, and odometry to simultaneously estimate planar vehicle motion and sparse scene structure. The authors extended this work[5] to recognize previous locations and to close the loop when a previous location is visited. A major difference between Se, *et al.*'s

system and our own is that our system is able to estimate six degree of freedom motion without stereo or odometry, which is important for applications like micro air vehicles. These vehicles have six degree of freedom motion, do not have odometry, and may image scenes that are distant when compared to the stereo baseline possible on the vehicle.

2 Long-term motion estimation

2.1 Overview

Even if gross local errors can be eliminated, gradual long-term drift in shape-from-motion estimates is inevitable when the camera moves between nonoverlapping views of the environment. Limiting drift in these cases requires (1) recognizing from the images that a previously mapped location has been revisited, and (2) using the previous estimates of the revisited location and the new observations to reduce drift in the new estimate. The second of these problems has been extensively studied in the simultaneous localization and mapping (SLAM) community, whereas the first problem is unique to image-based motion estimation.

This section describes a proof of concept system that addresses these issues. Subsection 2.2 describes our baseline shape-from-motion system, which attempts to eliminate gross local errors in the estimated shape and motion. Subsections 2.3-2.6 describe how multiple instances of this baseline system are combined with the idea of archived "rollback" states to reduce drift over time.

2.2 Baseline system

We have implemented a baseline system that, like existing shape-from-motion systems, uses two steps to simultaneously recover the camera motion and scene structure.

The first step is sparse feature tracking, which establishes correspondences between selected two-dimensional image features over multiple images. For this step, we use our Smalls tracker, which attempts to remedy the problems we have observed over several years using Lucas-Kanade tracking for shape-from-motion. In brief:

1. Smalls replaces Lucas-Kanade's unconstrained two-dimensional search with a one-dimensional search constrained by the epipolar geometry relating the two most recent images. The epipolar geometry is estimated using SIFT keypoints and RANSAC.
2. Smalls finds an initial estimate for a feature's position on the epipolar line in the new image using nearby SIFT matches, rather than the Lucas-Kanade pyramid.

3. Smalls replaces Lucas-Kanade's heuristics for extracting and eliminating features with heuristics more suitable for long-term shape-from-motion.

This algorithm relies on SIFT to do much of the heavy lifting, but can track a feature long after the feature would no longer be extracted as a SIFT keypoint. A more detailed explanation of Smalls is given in [6], along with an experimental evaluation that describes its performance on several difficult sequences.

The second step finds estimates of the six-degree-of-freedom camera position and three-dimensional point positions that are consistent with these feature tracks. Our system uses the variable state dimension filter (VSDF), which also provides an estimate of the joint covariance of the last several camera positions and the three-dimensional feature positions.

Used together, Smalls and the VSDF produce a very small number of gross local errors in the estimated camera position, even for difficult sequences. However, the system can still be defeated by extremely poor texture or repeated patterns in the image sequence. We revisit this issue in Section 3 below.

2.3 System states

In the following subsections we explain how multiple instances of our baseline system are combined with the idea of archived "rollback" states to reduce drift over time.

An *active state* S_i describes the system after processing the most recent image, with index i. The active state contains a list of image indices, I. During the system's initial operation, I will be the set of sequential indices $0, \ldots, i$. But as explained below, I is not generally this sequential set.

The active state also contains an instance of the Smalls tracker. This tracker instance gives a history of sparse, two-dimensional feature locations found by tracking through the image sequence I. Similarly, the active state contains an instance of the variable state dimension filter that includes six degree of freedom position and error covariance estimates for the camera at the time of each image in I, and estimates of the three-dimensional point positions of the features that were tracked through the images in I. Last, the active state also contains the SIFT keypoints extracted in the most recent image.

The system also maintains a collection of *rollback states*, S_{r_0}, S_{r_1}, ..., that describe the system state at previous times. As described in the following subsections, maintaining these rollback states allows the system to revert to and extend a previous motion estimate if the camera revisits a previously mapped location. The structure of the rollback states is the same as the structure of the active state. In the hypothetical example of the system's operation in subsection 2.6 below, the rollback states are the active states for every third image of the sequence. In the experimental results in Section 3, the rollback states are the active state for every tenth image in a prefix of the sequence. More details are given in those sections.

2.4 Extending states

When a new image arrives, a new state S' can be generated by extending a previous state S in the obvious way:

- I' is constructed by appending the new image index to I.
- The tracker is copied from S into S' and extended by tracking into the new image.
- The VSDF is copied from S into S' and extended using the tracking data for the new image generated in the item above.
- Any features identified as outliers in the new image by the VSDF are eliminated from both the tracker and the VSDF, starting in the new image.
- SIFT keypoints are extracted in the image and included in S'.

This procedure can used to generate a candidate active state from the current active state or from an archived rollback state, as described below.

2.5 Operation

When a new image with index i arrives:

1. A candidate active state $S_{i-1,i}$ is created from the new image and S_{i-1}, as described in the previous subsection.
2. A cueing procedure, described below in this subsection, is used to select a set of candidate rollback states, with indices $c_0, \ldots, c_k$, from the full set of archived rollback states.
3. Candidate active states $S_{c_0,i}, \ldots, S_{c_k,i}$ are created from the new image and the candidate rollback states $S_{c_0}, \ldots, S_{c_k}$.
4. The candidate state from $S_{i-1,i}, S_{c_0,i}, \ldots, S_{c_k,i}$ with the smallest estimated camera translation covariance for image i is adopted as the new active state S_i. To determine the smaller of two covariances, the largest principal components of the covariances are compared.
5. The new state S_i may be recorded for future use as a rollback state.

The cueing procedure in step 2 first identifies the rollback states whose most recent camera translation covariance estimate is smaller than the camera translation covariance estimate for image i in the candidate new state $S_{i-1,i}$. As in step 4 above, the smaller of two covariances is determined by comparing the translation covariances' largest principal components. The rationale here is that extending rollback states whose most recent camera covariance is already larger than the current candidate's camera covariance is unlikely to result in a smaller covariance.

The cueing procedure then further prunes the surviving candidate states by thresholding on the number of SIFT matches between the most recent image in the rollback state and image i. The survivors are the rollback candidates.

2.6 Example

Figure 1 illustrates the system's operation. Initially, images 0-7 have been processed without extending any rollbacks states, producing eight states with image index sets, {0}, {0, 1}, ..., {0, 1, ..., 7}. When image 8 arrives, a tentative new active state $S_{7,8}$ is generated with images {0, 1, ..., 8} by extending the previous active state S_7 into image 8. In this example, every third state (states S_0, S_3, and S_6) has been recorded as a rollback state, as shown in the diagram.

Then suppose that the cueing procedure identifies S_3 as a candidate rollback state for image 8. A candidate active state $S_{3,8}$ with indices {0, 1, 2, 3, 8} is generated by extending S_3 into image 8. If the camera position covariance for image 8 in $S_{3,8}$ is smaller than that in $S_{7,8}$, state $S_{7,8}$ is pruned and $S_{3,8}$ is adopted as the new active state.

The situation after several more images have been processed is shown in Figure 1. In this example, a state generated by extending a rollback state is adopted as the new active state, when rollback states S_3, S_6, S_9, S_{12} are extended into images 8, 11, 14, and 17, respectively. The final active state that results, S_{20}, has been generated using images {0, ..., 6, 11, 12, 17, ... 20}.

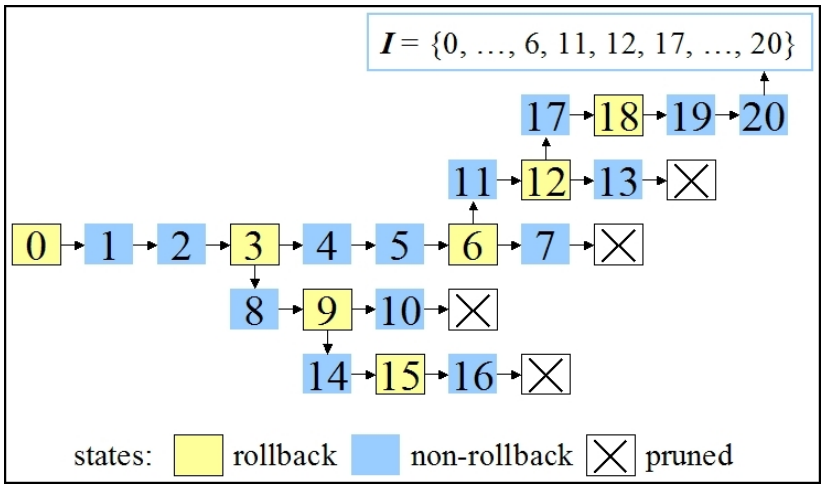

Fig. 1. A tree of states results from a sequence of rollbacks.

3 Results

3.1 Overview

This section briefly describes the system's operation on a 945 image sequence, in which the camera repeatedly moves between a number of nonoverlapping views of a complex, cluttered environment. Subsection 3.2 describes this sequence in more detail. Subsection 3.3 describes the system parameters, and subsection 3.4 describes the resulting motion estimates.

3.2 Input sequence

The image sequence is a 640×480, 945-image sequence taken from an IEEE 1394 camera with a wide field of view lens. The camera was mounted on a simple pushcart, and makes three forward and three backward passes through a cluttered scene. The camera motion is planar is x, y and contains rotation about the camera's y (image up) axis. But, it is important to note that the algorithm makes no assumptions that the motion is planar and computes six degree of freedom motion and three-dimensional structure estimates.

A few representative images from the first forward pass, which spans images 0 to 213, are shown in Figure 2. As shown in the images, the scene is highly non-planar, and the resulting images include severe occlusions, large image motions, and changes in overall intensity. Most areas are well-textured, but many parts of the image sequence contain poor texture (e.g., the saturated areas in images 2(b), 2(f), and 2(h)), repetitive texture, and one-dimensional texture (e.g., the overhead door in 2(h)).

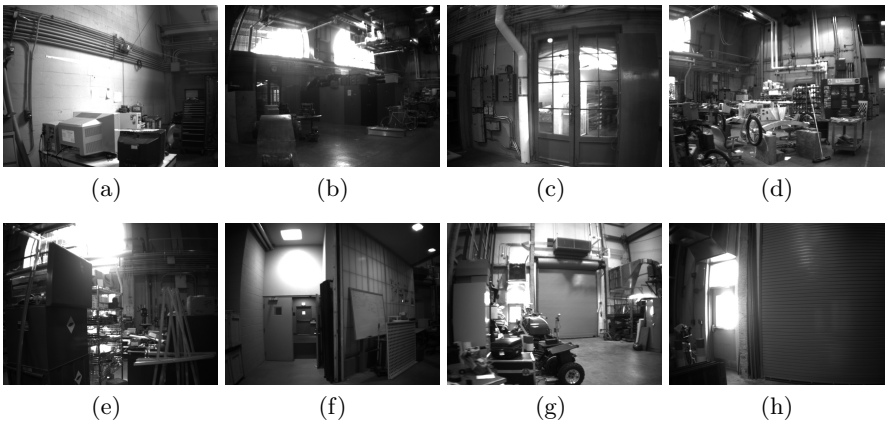

Fig. 2. Images 0, 43, 65, 110, 138, 167, 191, and 212 from the first forward pass.

The x, y camera translation during the first forward pass is illustrated in Figure 3. The estimated camera locations at the time of the images (a)-(h) shown in Figure 2 are marked. This illustration was generated from the system's estimates, which are described in subsection 3.4 below.

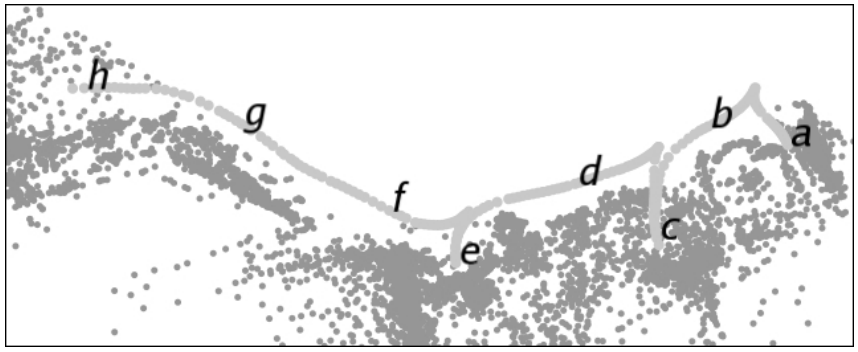

Fig. 3. An estimate of the camera motion and sparse scene structure during the first forward pass, generated by the proof of concept system. The estimated camera locations at the time of the images (a)-(h) in Figure 2 are marked in this figure. The estimates and the system parameters used to generate them are described in detail in subsections 3.3 and 3.4.

The motion during the five subsequent passes (three backward passes and two forward passes) traverses the same area, but may or may not include the visits to the side locations marked as "a," "c," and "e."

As shown in Figure 3, most of the imaged scene is on one side of the camera path. This is because the camera is mounted at a 45° angle, pointing forward and to the left of the cart. For shape-from-motion for ground vehicles, this viewpoint often provides a good combination of strong parallax to the side, which helps to improve short-term accuracy, and longer-term visibility of features in front of the vehicle.

3.3 System configuration

The system parameters that affect the system's performance most were chosen as follows.

- The number of VSDF initialization images and the number of images in the VSDF sliding window are both 10.
- The number of SIFT matches used for thresholding candidate rollback states in the cueing algorithm (subsection 2.5) is 300.
- The number of RANSAC trials for both the tracker's geometric mistracking step and the VSDF's mistracking detection step were 500.

- The tracker's geometric mistracking RANSAC threshold and the VSDF's mistracking RANSAC threshold were both 1.0 pixels.

As illustrated in Figure 1, the system can normally generate states by rolling back to and extending a state that was itself produced by a rollback, generating an arbitrary tree of states. In this example, however, the camera is known to map the entire area during the initial forward pass spanning the first 214 images of the sequence. So, for efficiency and for clarity of presentation here, rolling back was suppressed during the first 214 images, and in the remaining 721 images, rolling back was limited to states recorded during the initial, 214 image pass. As a result, the overall accuracy of the estimates during the 945 sequence is largely determined by accuracy of the recovered motion for the initial 214 images. It is important to note that no prior map of the first forward pass was used; the estimates during the first pass are those generated by the system during the first 214 images.

3.4 Estimates

During the first forward pass of 214 images, the system operates without rolling back and produces the camera motion and sparse scene structure estimates shown in Figure 3. In Figure 3, only the x, y components of the estimated camera translation and of the estimated scene structure are shown for clarity. However, the system does estimate full six degree of freedom motion and three-dimensional scene structure, including the three degree of freedom camera rotation, z camera translation, and z scene structure components, which are not shown in the figure.

Sufficient features are tracked throughout the entire initial pass to generate a qualitatively accurate map of the area. For example, the z translation components are extremely flat even though no assumption of planar camera motion was used in the estimation. One minor flaw, however, is the apparent growth in the global scale of the estimated translation and scene structure estimates late in the sequence, after point "f" late in Figure 3.

During the subsequent five passes, from image 214 to 944, the system is allowed to roll back to and extend states recorded during the initial pass. During these passes the system identifies a revisited location and extends the corresponding rollback state 25 times, as shown in Table 1. In the table, "$x \leftarrow y$" indicates that starting at image x, the rollback state corresponding to image y was reinstated and extended. Each rollback in the table corresponds to a correctly identified revisited location.

As indicated in the table, the system also fails to generate sufficient tracking data to estimate the camera's position three times, and the position is lost. These occur when the image is dominated by repetitive or one-dimensional texture. In these cases, the estimated camera translation covariance is set to infinity, so that the system recovers a position the first time the cueing algorithm finds 300 SIFT matches between the current image and a rollback image.

Table 1. The proof of concept system's reacquisition behavior for the 945 image sequence.

First backward pass:	228 ← 200, 278 ← 160, 301 ← 110,
(Images 214-380)	341 ← 060, 343 ← 070, 374 ← 040
Second forward pass:	454: position lost, 460 ← 180, 471 ← 190
(Images 381-493)	
Second backward pass:	494: position lost, 507 ← 200, 514 ← 190,
(Images 494-609)	551 ← 160 , 572 ← 110, 573 ← 120, 601 ← 040
Third forward pass:	678 ← 100, 726: position lost, 730 ← 180,
(Images 610-762)	742 ← 190, 753 ← 200
Third backward pass:	779 ← 190, 823 ← 150, 829 ← 120, 837 ← 110,
(Images 763-944)	872 ← 080, 907 ← 040, 934 ← 010

4 Conclusion

The system we have described is proof-of-concept, and three issues remain. First, our implementation is much slower than real-time. Second, the current policies for choosing states to archive as rollback states and for determining archived states to examine for each new image are simple, and do not bound the amount of space and computation required for these two tasks. To scale the system to much larger or infinite sequences, these policies must be improved. Third, the system is able to reduce drift in the current camera position by recognizing a previously mapped location, but does not use this information to refine previously estimated positions.

References

1. D. G. Lowe. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 60(2):91–110, 2004.
2. M. Montemerlo, S. Thrun, D. Koller, and B. Wegbreit. FastSLAM 2.0: An improved particle filtering algorithm for simultaneous localization and mapping that provably converges. In *Eighteenth International Joint Conference on Artificial Intelligence (IJCAI-2003)*, Acapulco, August 2003.
3. D. Nistér, O. Naroditsky, and J. Bergen. Visual odometry. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2004)*, volume 1, pages 652–659, Washington, D.C., June 2004.
4. S. Se, D. Lowe, and J. Little. Mobile robot localization and mapping with uncertainty using scale-invariant visual landmarks. *International Journal of Robotics Research*, 21(8):735–758, 2002.
5. S. Se, D. Lowe, and J. Little. Vision-based mapping with backward correction. In *IEEE/RSJ Conference on Intelligent Robots and Systems (IROS 2002)*, pages 153–158, Lausanne, Switzerland, October 2002.
6. D. Strelow. *Motion estimation from image and inertial measurements.* PhD thesis, Carnegie Mellon University, Pittsburgh, Pennsylvania, November 2004.