Structure and motion from SEM: a case study

Fred Nicolls

Department of Electrical Engineering, University of Cape Town
fnicolls@eng.uct.ac.za

Abstract

This paper describes an attempt to reconstruct a 3-D object from a set of 35 images captured using a scanning electron microscope. Point matching over overlapping triples of views is used to obtain an initial reconstruction, which is refined using bundle adjustment with the added knowledge that the sequence is closed. Intrinsic camera parameters are estimated via autocalibration under an affine assumption. Good results for the final metric reconstruction are obtained.

1. Introduction

Modern computer vision provides many techniques for reconstructing 3-D objects from uncalibrated sequences of images. This paper describes an attempt to solve the structure and motion problem for a set of images captured from a scanning electron microscope (SEM).

The test object used is a small aluminium block, of the order of 0.1mm in size, obtained by crudely milling away portions of a larger aluminium slab. To obtain multiple views, this object was mounted on a turntable and rotated in the view of the SEM. Figure 1 shows two frames from the sequence, which was 35 frames long in total. The images, each of dimension 1024×768, were taken at approximately equal angle increments of about 10 degrees. However, the motion of the turntable was neither precisely controlled nor monitored, and upon inspection it was evident that there was also no single axis of rotation.

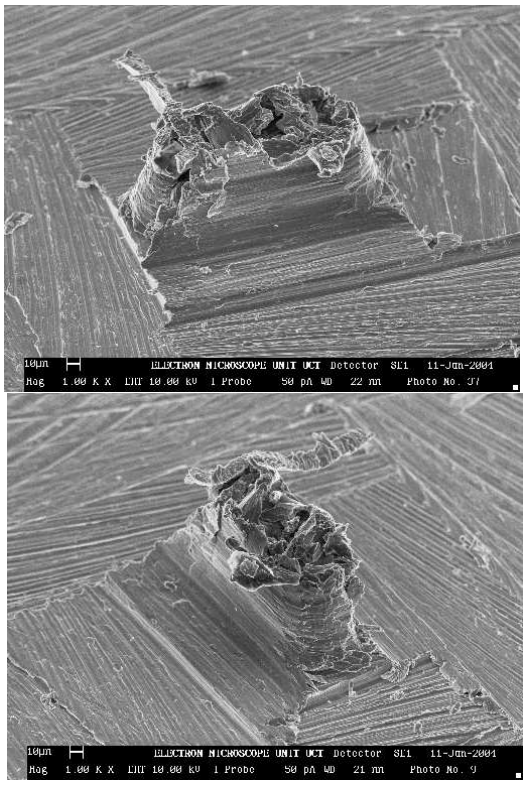


Figure 1: Two frames from the 35 frame input image sequence.

A scanning electron microscope operates under very different principles from optical imaging systems, and one cannot take it for granted that the assumptions made in computer vision will be appropriate. This issue is discussed in Section 2. From the outset, however, we make the assumption that a geometrical optics model is appropriate — if this is not the case then we expect to find inconsistencies in the application of the theory. In a sense, to obtain an accurate reconstruction is probably one of the best ways to validate that the assumptions of projective geometry are appropriate.

The approach taken in this work is to use standard feature-point based structure and motion techniques to obtain a projective reconstruction of the scene and the effective camera positions. This includes conventional outlier rejection, both for 2-view matches as well as for 3-view triplets. The only real complication in the reconstruction relates to the length of the sequence: there is no obvious way to combine the results from all 35 camera views into the reconstruction. Nonetheless, a simple method of reconstructing overlapping triples and stitching them together was adopted which, while not being without problems, provides a reasonably good reconstruction. A metric upgrade is then obtained by autocalibration. Details on all these methods are presented in Section 3.

The final result is a reconstruction of 3-D points on the surface of the object, as well as estimates of the locations of the (effective) cameras used to capture the images. These results are presented in Section 4.

2. Scanning electron microscopy

In scanning electron microscopy (SEM), a beam of electrons is used to form an image of a specimen. Since the SEM is a point-source (type 1) scanning microscope, at any time the illuminating beam is focused to a small spot on the object. This results in a signal which can be detected. The spot is scanned across the specimen and an image built up.

Depending on the configuration and the detectors used, the signal can provide information on a number of physical characteristics of the specimen, such as topography and atomic composition. In the secondary electron mode of operation, the beam results inelastic excitation of atoms to such high energy levels that electrons can overcome the work function and escape. These secondary electrons (which themselves have low energies $\leq$ 50eV) are then usually collected by a Everhart-Thornley detector. This detector consists of a positively-biased grid which attracts the secondaries, accelerates them onto a scintillator, and records the response. Figure 2 depicts the configuration for the beam at two different positions on the specimen.

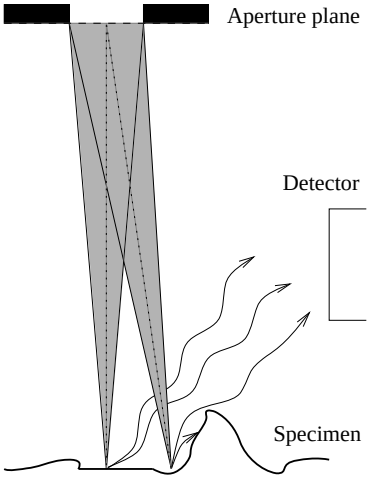


Figure 2: A SEM detecting secondary electrons.

It is not obvious that a SEM will produce an image that bears any resemblance to optical images — the mechanisms of image formation are entirely different. However, in [6, 5] it was demonstrated (for an autofocus application) that there is indeed some degree of equivalence: the notion of a point-spread function can be developed for a SEM, and is entirely analogous to that of conventional light optical systems. Furthermore the SEM conforms to the basic principles of geometric optics, although often with parameter values vastly different from those commonly encountered in optics. The apparent ease with which non-specialists interpret scanning electron micrographs bears testimony to this similarity.

Interestingly, with regard to photometric comparison one should regard the detector in SEM as being equivalent to a (possibly diffuse) light source in optical imagery, and the electron beam as being equivalent to the camera. This can be deduced though closer inspection of Figure 2: when the topography of the specimen causes an occlusion between the position of the beam on the specimen and the detector, the number of secondary electrons reaching the detector is decreased. This causes a reduction in the signal, which appears as a region of reduced intensity, or a shadow, in the resulting image.

3. Reconstruction method

The sequence of operations used to reconstruct the object from the image sequence is now described.

Good introductions to the theory and techniques of multiple view geometry can be found in [2] and [4]. The implementation "recipes" found in the latter are exceedingly useful during implementation. A fairly thorough exposition of how to perform reconstructions from indeterminate length video sequences is provided in [7]. A general review of autocalibration is presented in [3].

The process starts by applying the Harris corner detection to each of the images in the sequence. In total 1000 corners are extracted per image, each with a sufficiently high strength, and none of which are closer than 7 pixels to each other.

The corners from each image are then used to find point correspondences between each view and its neighbouring views, both forward and backward in the sequence. Block matching is used with a normalised correlation distance measure and blocks of size 15×15. With a minimum match value of 0.8 it was found that around 100 matches were found for each of the images in the sequence.

The fundamental matrix linking adjacent views is estimated using RANSAC on the correspondences. The images are subsequently rectified, and a guided matching stage performed. This increases the number of good correspondences from 100 to around 500. The minimum match value used here is the same as for the initial matching stage.

Since for a single set of corners the matching was performed both forwards and backwards, we effectively have correspondences over tracks of length three spanning triplets of images. Robust estimates of the trifocal tensor linking all sets of three adjacent views are then made, using RANSAC with an inlier distance of 1.25 pixels. It was found that about 300 matches per triplet survive this operation.

The trifocal tensor provides a strong constraint, and it is unlikely that any of the surviving matches will be incorrect. The triples therefore provide excellent initial values

for a complete reconstruction process.

At this stage a number of options can be exercised — each triplet can be used independently to provide a reconstruction of the points in the scene, but no uniformly best method appears to exist for how to combine these separate reconstructions. In this work a simple method of extending by resectioning was adopted. Since adjacent triples overlap by two views, a reconstruction from the first triple can be extended to the next simply by estimating the final camera of the new triple in the projective coordinate system of the first. That is, for a new triple the two known cameras can be used to estimate the world locations of the next set of points, from which a world-to-image transformation can be obtained corresponding to the new camera. This process of resectioning can be continued until the entire sequence has been reconstructed.

A problem with this approach is that the reconstruction is subject to drift, as there is no mechanism to prevent a gradual accumulation of errors as views are added. The effect of this drift would decrease if points were tracked for longer than just over image triples, but this would require the development of an affine (or projective) invariant match procedure. Further details relating to drift in reconstructions can be found in [1].

A simple method to reduce this drift makes use of the fact that the sequence is closed, with the first image being the natural neighbour to the last. It is easy to include this information into a bundle adjustment stage which uses the reconstruction described previously as an initial estimate of the solution. With 35 views containing around 12000 points in total this is a large problem, and consumes a large amount of computer memory. Since the observation matrix is banded and diagonal-dominant some reduction in computation could be achieved if desired. Alternatively we could just use a subset of the available points — there are far more than are needed for an accurate reconstruction of structure and motion in any case.

The final outcome of the procedure described is a projective reconstruction of points on the surface of the object, as well as locations of the cameras used to generate the images. This reconstruction differs from a Euclidean one by an unknown homography. To upgrade the reconstruction to metric, we need some information about the effective camera parameters used during the image capture. Since for the scanning electron microscope we have no means of explicitly calibrating either the intrinsic or extrinsic camera parameters, autocalibration at this point is essential.

The microscope settings were not changed while the object was rotated in the view, so an assumption of a fixed camera is appropriate. Also, it being a precise piece of equipment there is every reason to believe that the pixel skew can be assumed to be zero, and the pixels square. It is important that this latter assumption be made — turntable motion represents a critical motion sequence (CMS) in multiple view geometry, and without constraints a projective ambiguity arises in the component of the reconstruction in the direction of the screw axis [4, p.492].

One final assumption that is made is that the SEM can be represented as an affine imaging system, so the last row of each camera matrix is [0 0 0 1]. This assumption was made explicit when it was found that a full projective autocalibration produced unstable results when applied to the problem. It is believed that this instability is caused by near singularities in the projective autocalibration procedure when the actual camera geometry tends to be affine, although no attempt was made to validate this claim. Although not an assumption that we wanted to make at the outset, it had been observed throughout this project that the SEM imaging system did indeed seem to be affine.

An affine camera can be decomposed as

$$P_A = \begin{pmatrix} \alpha_x & s & 0 \\ 0 & \alpha_y & 0 \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} \hat{\mathbf{R}} & \mathbf{t} \\ \mathbf{0}^T & 1 \end{pmatrix},$$

where the leading matrix in this decomposition contains the intrinsic camera parameters. This decomposition forms the basis of an autocalibration procedure proposed by Quan [8], which was used in this work. In short, an upper triangular matrix $\mathbf{Z}$ is found which, when applied to all the cameras, yields a decomposition that satisfies the constraints on the intrinsic parameters. The elements of $\mathbf{Z}$ are found by a nonlinear least squares procedure, initialised by the identity matrix.

4. Results

It is difficult to present a reconstruction of a point cloud in print — it is useful to be able to navigate around it (with for example a VRML viewer) to convince oneself that the reconstruction is accurate. Nonetheless, Figure 3 shows a reconstruction of the structure and motion of the aluminium block sequence *before* making use of the assumption that the sequence is closed. The lines in the figure indicate the principal rays of the estimated cameras. Since these cameras are assumed affine they effectively lie on the plane at infinity. Close inspection reveals that the reconstruction is not sharp, and that points corresponding to the same image features in different triplets appear as distinct and different. The RMS reprojection error for this reconstruction is 0.3204 pixels.

If we make use of the fact that the sequence is closed, with the first image following the last, then visually better results are obtained. Figure 4 shows reconstructions from roughly equivalent viewpoints both before and after including this assumption. It is clear that fine structures are reconstructed more accurately and with less spread in the second case. The RMS reprojection error for the closed sequence reconstruction is 0.3252 pixels.

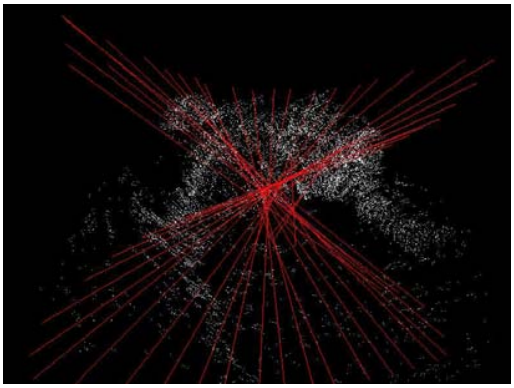
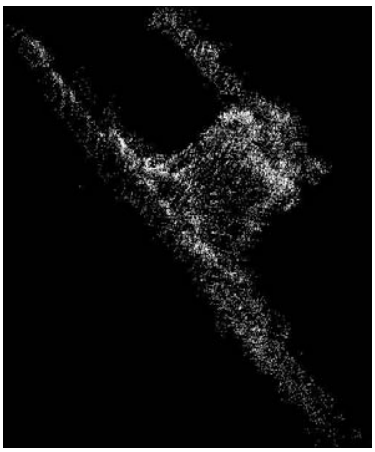


Figure 3: Reconstruction of points and camera views without the assumption that the sequence is closed.

(a) Without closed sequence assumption.



(b) With closed sequence assumption.

Figure 4: Effect on the reconstruction of using the assumption that the sequence is closed.

5. Conclusion

The research presented in this paper is at an early stage, and the number of improvements that could be made are almost too many to number. The most immediate improvement would be obtained by tracking the points over longer durations. This would probably necessitate the development of an affine (or projective) invariant feature matching procedure.

The assumption of the SEM being an affine imaging system appears to be valid and accurate. In retrospect this may have been a reasonable assumption to make at the outset: from a geometrical point of view the field-of-view of the images is around 1mm, while the working distance is 22mm. This corresponds a deviation from parallel projection by of the order of only about 1°. At higher magnifications the deviation may be expected to reduce even further.

6. References

[1] K. Cornelis, F. Verbiest, and L. Van Gool. Drift detection and removal for sequential structure from motion algorithms. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 26(10):1249–1259, October 2004.

[2] David A. Forsyth and Jean Ponce. *Computer Vision: A Modern Approach*. Prentice Hall, 2003.

[3] A. Fusiello. Uncalibrated Euclidean reconstruction: a review. *Image and Vision Computing*, 18:555–563, 2000.

[4] Richard Hartley and Andrew Zisserman. *Multiple view geometry in computer vision*. Cambridge University Press, second edition, 2003.

[5] F. C. Nicolls, G. de Jager, and B. T. Sewell. Use of a general imaging model to achieve predictive autofocus in the scanning electron microscope. *Ultramicroscopy*, 69:25–37, August 1997.

[6] Frederick Nicolls. The development of a predictive autofocus algorithm using a general image formation model. Master's thesis, University of Cape Town, Rondebosch 7700, South Africa, February 1996. http://www.dip.ee.uct.ac.za/∼nicolls.

[7] Marc Pollefeys, Luc Van Gool, Maarten Vergauwen, Frank Verbiest, Kurt Cornelis, Jan Tops, and Reinhard Koch. Visual modeling with a hand-held camera. *International Journal of Computer Vision*, 59(3):207–232, September 2004.

[8] Long Quan. Self-calibration of an affine camera from multiple views. *International Journal of Computer Vision*, 19(1):93–105, May 1996.

7. Acknowledgements

The author wishes to thank Trevor Sewell from the UCT Electron Microscopy Unit for capturing the SEM dataset used in this work.