Portable VXL System for Computing Structure From Motion

David McKinnon*, Kurt Kubik and Brian Lovell
Intelligent Real-Time Imaging and Sensing (IRIS) Group
The School of Computer Science and Electrical Engineering
The University of Queensland, Australia QLD 4072
Contact Author* : s341552@student.uq.edu.au

Abstract

This paper considers the recovery of the epipolar geometry for the case of calibrated/uncalibrated two view relations. Two view relations are the basis for accurate calibration of stereo rigs and are often used in the solution to sequential processing of digital image streams to recover the surface structure of a scene or otherwise known as, structure from motion (SFM). The outcomes of this research are a freely available implementation of the algorithms required to determine an accurate solution to the epipolar geometry of a two-view relation.

1. Introduction

The recovery of the epipolar geometry for a pair of views is a well known process. As early as the 1869 the projective geometry of the epipolar relation had been recognised, first by Sturm[11] and more recently by Longuett-Higgins[7] and Faugeras[2]. The computation of the SFM of a pair of images is computationally demanding. The system discussed in this paper has been written in the VXL C++ computer vision libraries (www.robots.ox.ac.uk/~vxl), which supply an excellent basis for computer vision research and development. The algorithms required to compute the SFM of an image sequence are publically available in VXL's Multiple View Geometry Library (MVL), and executables for the SFM program are available for Windows and Linux architectures (www.csee.uq.edu.au/~iris).

In practice the problem of the recovery of the surface structure from the scene involves the solution to the epipolar relation between the cameras (this epipolar configuration is demonstrated in Figure 1), namely the rotation of the optical axes from one camera to the next (R), and the translation of the optical centers of the cameras through 3D world space (t). The P (or camera) matrix (1) is responsible for the projection of the 3D projective co-ordinates $X = (x, y, z, 1)$ to the 2D projective image co-ordinates $x = (u, v, 1)$;

$$x_i = PX_i = A[R \mid -Rt]X_i \tag{1}$$

A represents the intrinsic properties of the camera conveyed by the pinhole model of the camera [15]. Lense effects such as radial distortion and aberations are not considered in this model of the camera. The matrix A is composed as follows,

$$A = \begin{pmatrix} \alpha_x & s & x_o \\ 0 & \alpha_y & y_o \\ 0 & 0 & 1 \end{pmatrix}$$

These intrinsic parameters represent the aspect ratio, focal length (α_x, α_y), image center (x_o, y_o) and skew (s) of the camera. This is a necessary consideration when mapping the light hitting the camera's lense as it is projected through the camera center to the CCD array or exposure. To the solve the epipolar relation there are two generalised approaches. Depending upon the practicality of the situation to which they are being applied, they can be used at one's discretion.

The first case is that of the calibrated camera. This requires prior knowledge of the camera/s and publically available programs such as Zhengyou Zhang's *'Easy Calib'*[16]. Once the camera intrinsics are known, the projective equations (1) can be determined up to a Euclidean basis after the solution of the Longuett-Higgins equation (2) which in turn allows a determination of the camera matrices. Note the x' denotes a point in the second image.

$$x'^T Ex = 0 \; where \; E = t \times R \tag{2}$$

The problem formulation in the calibrated case is impaired by the fact that it restricts the determination of changes in the cameras intrinsics such as focal length. This means that the auto-focus feature on most modern digital cameras must be disabled and the focal length must remain static after calibration.

The second more general case of camera calibration is that of the uncalibrated camera. This case is the focus of

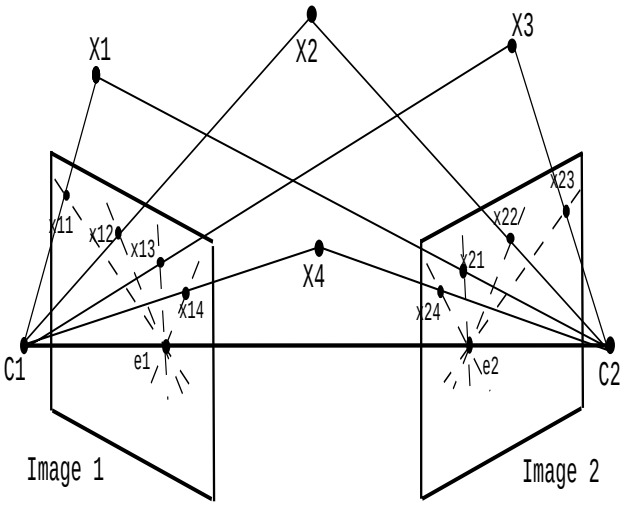

Figure 1. Epipolar relationship between points visible in two views.

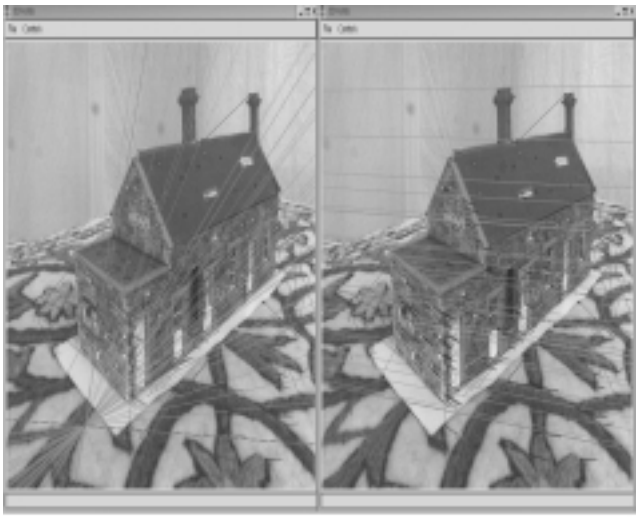

Figure 2. (a) All matched candidates with estimated epipolar geometry (b) Inliers, with estimate of epipolar geometry

most contemporary structure recovery algorithms, and has received the majority of the attention in the modern literature. The important advantage of the uncalibrated case is that the structure recovery algorithm does not need access to the camera prior to processing the sequence to recover the intrinsic properties of the camera. Features of the camera such as the auto-focus will not affect the reconstruction as estimates of the cameras intrinsics can be calculated for each frame in the pair or sequence.

The discovery that allows the computation of the structure in the uncalibrated case was Faugeras's F matrix. This is essentially a generalisation of the Longuett-Higgins equation for the case of unknown camera intrinsics.

$$x'^T F x = 0 \; where \; F = A'^{-T} t \times R A^{-1} \qquad (3)$$

Clearly the complexity of the structure recovery process is increased with the addition of more unknowns in the camera intrinsic matrices. But it is possible to recover solutions to different camera bases (Affine, Similarity & Euclidean) from several of these uncalibrated relations [9, 3].

So given an F matrix the Projective camera matrices P and P' can be derived as;

$$P = [I \mid 0] \; and \; P' = [e' \times F + e' v^T \mid \lambda e'] \qquad (4)$$

where e' is the nullspace of F^T (or the epipole in the second image), v is the vector for the plane-at-infinity which is the plane containing the vanishing points of parallel lines in the scene. λ is an arbitrary constant that merely changes the changes the scale of the resulting reconstruction. The most challenging component in recoverying the camera matrices in the uncalibrated two-view case is solving for the plane-at-infinity [4]. Note that given a proper estimate of the plane-at-infinity the resulting reconstruction will be Affine, where as other estimates will result in a Projective reconstruction.

Gaining a reconstruction from the E matrix requires a slightly different method [5]. The E and F matrices will be refered to as C from this point.

2. Background

This section of the paper provides an overview of the theory maintained in the solution to the un/calibrated camera matrices.

2.1. Robust Determination of Scene Structure

In order to solve C, features must be recovered from both images and putative matching candidates isolated [17]. This is a relatively simple process requiring an initial feature extraction from both images followed by some process to find matches between features in each image. This paper uses corners (Harris corner detector) as the feature to be matched across the images and NCC template matching to find the putative correspondences. It is also possible to use lines or curves exclusively or in conjunction with corners to potentially increase the accuracy of the two-view relation. Once putative matching candidates have been isolated in both images, proceeding in a Least-Squares fashion results in a solution for C. But as demonstrated in Figure 2. the presence of incorrectly matched candidates coupled

with errors in corner localisation will degrade the accuracy of the resulting relation. For close range scene modelling this can be catastrophic for the resulting accuracy of the re-projected 3D point estimates. In order to cater for these errors, robust methods of weighting the validity of each of the corner-to-corner candidates are employed.

There are many different ways to perform robust weighting of matched candidates. This work implements the use of three methods that have enjoyed the most attention in the literature. These are RANSAC [12] (Random Sampling Concensus), MLESAC [14](Maximum Likelehood Sampling Concensus) and LMedSq [17](Least Median of Squares). These differing methods all use Monte-Carlo bucketing to gather 7-point subsamples from the entire set of matched candidates, where 7 corner-to-corner matches are the minimum number required to solve for C. These 7-point subsamples can generate a hypothesis for the relation governing all matched candidates, through the generation of multiple hypotheses and testing of the fit of each hypothesis a 7-point basis for the relation is determined. This basis and its corresponding estimate of C will either maximise the number of correctly matched corner-to-corner candidates or inliers (RANSAC) or minimise the overall residuals of the fit (MLESAC & LMedSq).

The robust cost function in each case is expressed as :

$$-RANSAC \; : \begin{cases} d < threshold, & inlier \\ d > threshold, & outlier \end{cases}$$

maximises the number of inliers, where d in this case is the following epipolar error measure,

$$d = \sqrt{(x'^T C x)^2 \; + \; (x^T C^T x')^2}$$

$$-MLESAC \; : \begin{cases} \frac{d}{\alpha} < \chi^2, & inlier \\ \frac{d}{\alpha} > \chi^2, & outlier \; \& \; \frac{d}{\alpha} = \chi^2 \end{cases}$$

MLESAC proceeds by finding the 7 point basis that minimises the combined error of all the matched correspondences, where outliers are capped to the threshold used for comparison.

$-LMedSq$: seeks to minimise the median error d of the resulting vector of errors from each of the 7-point bases being tested. Thus finding the basis resulting in the minimum median residual error.

2.2. Non-Linear Minimisation

Once an initial estimate for the relation has been determined, the inliers of the matched corner-to-corner candidates can be used to further increase the accuracy of the estimate for C by a gradient descent type algorithm (this work uses the Levenburg-Marquardt algorithm).

This work has implemented two Non-Linear minimisations, one uses a rank-2 parametrisation of the C matrix which is not available for the straight Least-Squares fitting procedure discussed in the previous subsection [17]. The other parametrisation was suggested by [14], and looks to augment the given 7-point basis for the relation to arrive at an optimal estimate for all the inliers.

The process of Non-Linear minimisation is potentially flawed if an adequate initial estimate for the relation is not provided. The error surface has local minima that can distract the gradient descent process from the true global minima.

2.3. Triangulation

Once the C relation has been determined for a pair of views and the initial set of camera matrices constructed, it is still necessary to back-project the 2D points from both images to their mutual position in 3D space. In the calibrated case this process will result in the 3D Euclidean points or sometimes called a metric set of scene points. In the uncalibrated case, the process of recovering the structure suffers from the indeterminacy of the cameras intrinsics properties thus rendering only a Projective or Affine solution to the structure tractable.

In either case the process of the back-projecting the rays to their point of intersection is generic. Again there are range of possible methods to perform this back-projection [10].

2.4. Degenerate Point Configurations

To further complicate the process of reconstructing a scene from a set of features correspondences, there are certain configurations of points and/or camera motions that defy the usual epipolar description. These point configurations are termed degenerate.

Degenerate configurations, if not detected, will lead to spurious two-view relations that can break down image sequence processing or render a pair of views unreconstructable. Additionally, point sets that lie close to degenerate configurations can also exhibit poor estimates for the resulting two-view relation.

Point configurations that will lead to degeneracies are generalised in a practical sense to;

- Cyclo-rotation, where the only motion between frames is a rotation of the camera about its optical center.

-Movement of the camera relative to a plane consuming the entire field of vision, or even still a dominant plane in the scene that attracts all the feature matches.

-There are also classes of surfaces that are degenerate these are called *critical surfaces* [8]. Point pairs that are matched on these surfaces will render portions of the matched candidates degenerate. Critical surfaces include cones, cylinders and hyperbolic paraboloids.

To overcome these practical anomalies, the work of Torr [13, 12] describes a process of robust model fitting, called GRIC (Geometrically Robust Information Criteria) following the work of Kanatani [6]. This process finds both a homography (non-degenerate two-view relation) and a C matrix for a pair of views and compares the MLE of the fit in both cases, chosing the model that minimises the GRIC score for the views in question. The homography calculated is a 2D projective transform between images that has no degeneracies but doesn't encode the epipolar geometry for the views. This means that you can maintain point tracks through a degenerate pair but you are unable to reproject the point pairs to gain structure.

In the case of just two views this can be utilised as a flag to indicate that the image pair is invalid, or for sequence processing this will save a breakdown in the tracking process.

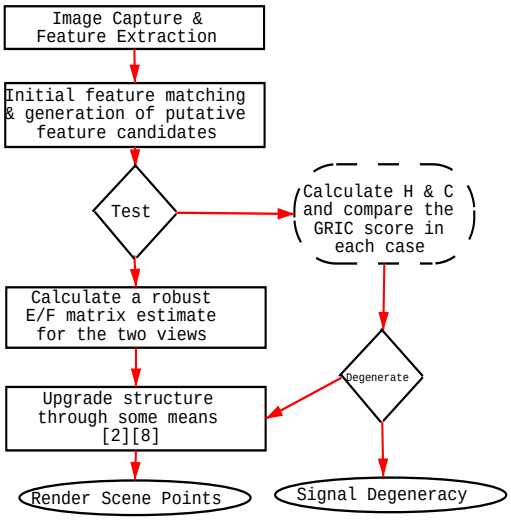

Figure 3. Block diagram of the two-view reconstruction process

3. Implementation

The theory discussed in the previous section has been coded in C++ using the open source VXL libraries and is freely available for download from the Multiple View Geometry (MVL) library contained within the OXL (Oxford) library in the public distribution (www.robots.ox.ac.uk/~vxl).

The VXL libraries provide an ideal environment for C++ implementations of computer vision and image processing type code. The libraries offer support for most common image formats (including support for very large images). In addition VXL provides most common image processing functions (filters, feature extraction, etc.) and an excellent numerical library which maintains C++ wrappers for a large collection of publically available NETLIB (www.netlib.org) FORTRAN code.

Perhaps the most impressive faecet of the VXL library collection is it's cross-platform capabilities. By abstracting all the machine/OS dependent components of the library such as GUI creation and standard C++ library calls VXL can operate on all the most common operating systems (excluding Macintosh & BEOS). There is a full collection of STL containers and capabilities to produce .mat files compatable with MATLAB (or OCTAVE) mathematical environments.

A thoughtfully coded C++ library such as VXL gives the programmer capabilities for semi-rapid development of computer vision applications that enjoy the full benefit of C++ compiler optimisation and cross-platform portability.

3.1. Block Diagram of the Two-View Structure Recovery System

Figure 3. is a block diagram of the two view structure recovery system discussed in the background section.

3.2. 3DWorlds

The application that utilises the routines discussed above is called 3DWorlds. 3DWorlds requires an image sequence in the following format image.###.format where .format includes .pgm, .ppm, .jpg, .tif, .mit, .viff and .rgb image types. Figure 4 shows the project options available for 3DWorlds, where the results of the estimates of the 2D feature locations, H/C matrices and corresponding 3D points are written to text database files where they can be analysed by other means.

The results of 3DWorlds for known motion synthetic image sequences has validated the accuracy of the routines for the case of medium baseline SFM sequences, with wide-baselines and small baselines currently being beyond the capabilities of the feature tracker (NCC template matching).

Figure 5 shows the results of the routines on a couple of two-view pairs where zooming of the camera along its optical center is present and straight stereo motion of the camera (only the second image of the pairs are displayed). Note that in the testing of the GRIC algorithm, all known degeneracies where detected with no false detections, though sequences with very small-baselines are flagged as degenerate, reflecting their poor encoding of structure.

The 3D points generated by 3DWorlds can either be of a Projective or Quasi-Euclidean basis [1]. In the case of

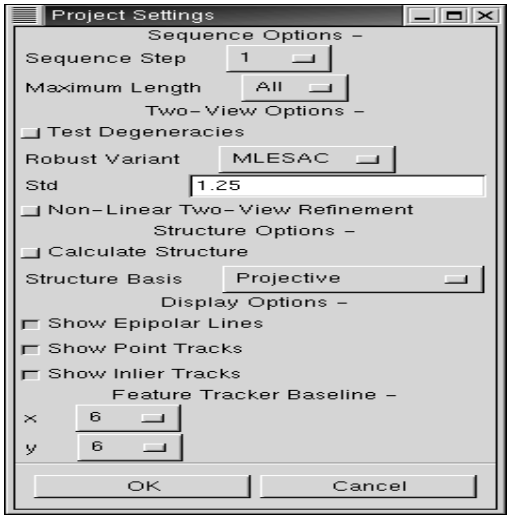

Figure 4. 3DWorlds project options

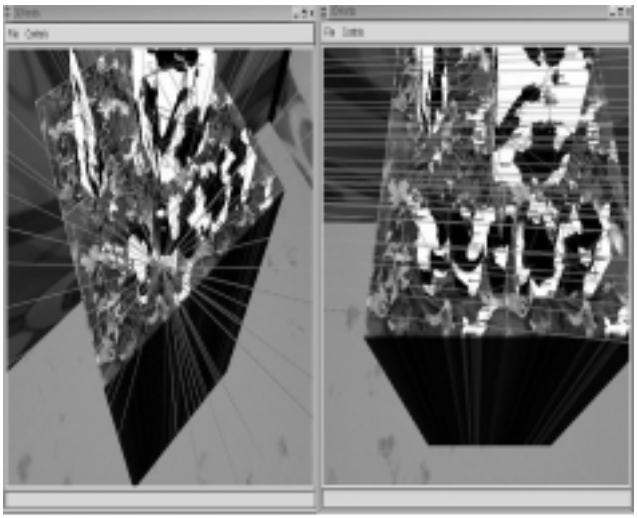

Figure 5. (a) Synthetic image with epipole at center (b) Sythetic stereo image with epipole at infinity

the latter the 3D scene points reflect more closely the actual 3D structure of the scene, though falling short of a proper Euclidean representation.

4. Conclusions

The concepts related to creating a sparse reconstruction of a surface seen in two views from a camera have been discussed. The system presented is available for use in other multiple-view geometry related research, and a executable program that gives a database of text file estimates of homographies and F matrices between image pairs. Along with camera matrices depicting either Projective or Quasi-Euclidean reconstructions of the scene with associated 3D point estimates.

Ideally the system should also incorporate the full Stratified recovery of a Euclidean reconstruction, by performing a self-calibration procedure [9] that will upgrade the Affine/Projective reconstruction to a Euclidean reconstruction. This is a topic to be considered in future work.

References

[1] P. Beardsley, A. Zisserman, and D. Murray. Sequential updating of projective and affine structure from motion. *International Journal of Computer Vision*, 23(3):235–259, 1997.

[2] O. Faugeras. What can be seen in three dimensions with an uncalibrated stereo rig? In *ECCV92*, pages 563–578, 1992.

[3] A. Fusiello. Uncalibrated euclidean reconstruction: A review. *Image and Vision Computing*, 18(6-7):555–563, 2000.

[4] R. Hartley. Cheirality invariants. In *DARPA93*, pages 745–753, 1993.

[5] R. Hartley and A. Zisserman. Multiple view geometry. Cambridge University Press, 2000.

[6] K. ichi Kanatani. Automatic singularity test for motion analysis by an information criterion. In *ECCV (1)*, pages 697–708, 1996.

[7] H. Longuet-Higgins. A computer algorithm for reconstruction of a scene from two projections. *Nature*, 293:133–135, 1981.

[8] S. Maybank. Theory of reconstruction from image motion, springer. *BERLIN*, 93:1992, 1992.

[9] M. Pollefeys. Self-calibration and metric 3d reconstruction from uncalibrated image sequences. In *PhD thesis*. Katholieke Universiteit Leuven, Belgium, 1999.

[10] C. Rothwell, O. Faugeras, and G. Csurka. A comparison of projective reconstruction methods for pairs of views. *Computer Vision and Image Understanding: CVIU*, 68(1):37–58, 1997.

[11] R. Sturm. Das problem der projectivit t und seine anwendung auf die ae chen zweiten grades. *Math. Ann.*, 1:533–574, 1869.

[12] P. Torr. An assessment of information criteria for motion model selection. *Proc. IEEE Conf. Computer Vision and Pattern Recogn*, pages 47–53, 1997.

[13] P. Torr, A. Fitzgibbon, and A. Zisserman. The problem of degeneracy in structure and motion recovery from uncalibrated image sequences. *Int. J. Computer Vision*, 32(1):27–44, 1999.

[14] P. Torr and A. Zisserman. Mlesac: A new robust estimator with application to estimating image geometry. *Computer Vision Image Understand.*, 78:138–156, 2000.

[15] R. Tsai. A versatile camera calibration technique for high-accuracy 3-D machine vision metrology using off-the-shelf TV cameras and lenses. In L. Wolff, S. Shafer, and G. Healey, editors, *Radiometry – (Physics-Based Vision)*. Jones and Bartlett, 1992.

[16] Z. Zhang. A flexible new technique for camera calibration. In *Technical Report MSR-TR98 -71*. Microsoft Research, 1999.

[17] Z. Zhang, R. Deriche, O. Faugeras, and Q.-T. Luong. A robust technique for matching two uncalibrated images through the recovery of the unknown epipolar geometry,. *Artificial Intelligence, December 1995*, 78:87–119, 1995.