Self-calibration of a stereo rig in a planar scene by data combination

Joss Knight and Ian Reid
Department of Engineering Science
University of Oxford
Oxford OX1 3PJ, UK
[joss,ian]@robots.ox.ac.uk

Abstract

We present a very simple and effective method for eliminating the degeneracy inherent in a planar scene, and demonstrate its performance in a useful application — that of binocular self-calibration.

The projective geometry of planar scenes suffers from a two-fold ambiguity: projective structure cannot be recovered from a pair of images of the scene, and transformations in projective 3-space are underconstrained for features lying on a plane. We overcome both these problems by combining feature data from different pairs of images where we know the relationship between the images to be identical. In essence this generates images of a non-degenerate scene which can then be processed using standard algorithms. The only constraint is that the binocular head must be able to make repeated identical rotations about its axes.

The benefits of this technique are demonstrated by its use in solving the problem of self-calibration. The use of data combination enables standard general scene stereo calibration algorithms to be used. Simulation and real scene tests prove the validity of the technique, and show it to be a worthwhile addition to the limited planar calibration literature.

1. Introduction

We have recently been interested in the automatic calibration of our stereo rig, Yorick (Figure 1). Yorick is a pan-tilt head mounted on a mobile robot GTI, which together are used in visual navigation applications. One of the prerequisites for use of Yorick in such applications is that we obtain at least an estimate of the internal parameters of both cameras.

Most of the previous work in affine or metric self-calibration of stereo rigs, [1–5, 12], is predicated on the ability to compute projective structure and/or the locations of the cameras by viewing a general scene. We look at the common case of a scene dominated by a single plane, for which the fundamental matrix relating two images of the scene is underdetermined, and thus calculation of projective structure is not ordinarily possible.

There has been limited previous work in self-calibration in these degenerate circumstances. Most assumes various scene constraints such as known metric structure, for example [9, 11]. Most notably, Triggs [10] has shown that calibration is possible from 5 or more monocular views of an unknown planar scene. The disadvantage of Triggs’ method is that it relies on a bundle adjustment procedure with no clear-cut initialisation. [6] by contrast uses geometry for stereo calibration, relying on the head’s ability to make zero-pitch screw motions about its axes (that is, rotations about an axis with no translation along the axis), and with three stereo views can recover estimates of the camera focal lengths, and the plane at infinity, assuming all other camera parameters are known. In this paper instead we use the head’s ability to make repeated identical rotations about an axis (possible, for instance, if the head possesses odometry, or it is driven by stepper motors). This enables us to use data combination to generate a virtual scene containing multiple planes and therefore no degeneracy. This scene can be processed with standard calibration algorithms and thus all the camera parameters can be recovered.

The idea of data combination is not new, but rarely ex-
plicitly defined. Its use in resolving scene degeneracy is not inextricably linked to self-calibration, and in fact is useful wherever 3D geometric relationships are being made for a degenerate scene, such as in structure recovery, or navigation. Man-made environments tend to contain a large number of planes, and as a result degeneracy is an unavoidable problem.

In section 2 we give the mathematical notation used in the paper, and summarise the general scene calibration method we use once problems of scene degeneracy have been resolved. §3 explains why resolving this degeneracy is necessary. In §4 we explain our method of data combination, and in §5 we summarise our algorithm. §6 gives performance results for tests of the algorithm. We make conclusions in §7.

2. Preliminaries

We use homogeneous 3-vectors to represent image points and lines, written in lower-case bold (x, l); homogeneous 4-vectors to represent scene points and planes, written in upper-case bold (X, Π); and matrices are written in upper case teletype style (H, Π).

2.1. Non-degenerate case calibration

Let us visualise points in ordinary, euclidean space X undergoing some rigid motion to points X′, denoted by the 4 × 4 matrix D, such that X′ = DX.

If we have stereo views of a general scene (non-planar) we can calculate projective structure, which is related to its euclidean counterpart by some (usually unknown) 4 × 4 projective to euclidean update matrix, denoted H_{PE}; i.e. projective points X_p are related to euclidean points by X = H_{PE}X_p. Given these equations we can write down the relationship between projective structure before and after the motion, T:

\[ T = H_{PE}^{-1}DH_{PE}. \]  

H_{PE} contains the camera internal parameters, and the location of the plane at infinity. [3–5, 12] discuss methods of factoring T to get at these parameters. They show that, if no scene or calibration constraints are applied, two T matrices representing motions about different, non-parallel axes are required before H_{PE} can be recovered uniquely (where H_{PE} must be identical for each motion, meaning the camera calibrations do not change, and the inter-camera relationship must be fixed).

In this paper, we recover one T matrix representing a rotation about the elevation axis, and one representing a rotation about the pan axis; then we mutually decompose these to give H_{PE} (and hence the camera calibration), and two Π matrices which provide the euclidean locations of the head axes. This is essentially an identical procedure to that described in [5], except that because the scene is planar, we require further views to resolve the degeneracy. In addition we can use the planarity constraint to improve the results.

3. Planar Degeneracy

We assume some familiarity here with epipolar geometry. In summary, for two cameras viewing a general scene, for a point in one image there will be a line of possible correspondences in the other, since an image point is the projection of any scene point along its ray. The relationship is encapsulated in the fundamental matrix F, and F can be use to calculate projective camera matrices for the cameras, that is, matrices which can be used to backproject matched features to provide projective structure.

However, if the scene is planar there is only one scene point associated with each image point. This means there is a unique point-to-point mapping between two images of the scene, encapsulated by a homography H, and the fundamental matrix is underdetermined regardless of the number of point matches on the scene plane. Hence projective structure cannot be calculated.

3.1. Ambiguity of 3D transformations

Normally the matrix T of equation 1 can be determined from sufficient feature correspondences in projective 3-space, using the relationship X'_p = TX_p for points in projective space before and after the motion, X_p and X'_p. If the points X_p are constrained to lie on plane Π, then \( \Pi^\top X_p = 0 \). Hence it follows that for any point in 3-space not on the plane, V:

\[ (T + VΠ^\top)X_p = TX_p + V0 = TX_p = X'_p. \]

This is a four-parameter ambiguity because the scale of V matters, which implies we require at least two matched off-plane points to calculate T (this is unrelated to the similar result for fundamental matrices).

4. Data Combination

Figure 2(a) shows a stereo pair of cameras viewing a planar scene before and after some motion. Because the transformation that takes the left camera’s coordinate system to the right’s is fixed during the motion, it is equally valid to visualise the motion as being that of the scene rather than the cameras. The result is we have a single stereo view of two planes as in Figure 2(b). Now the scene being viewed is no longer degenerate, and its epipolar geometry can therefore be calculated.

In practice we combine features detected in the two left camera images, and right camera images, as if they had
originated from the same image. Clearly the quality of projective structure that results from calculating the epipolar geometry (and hence projective camera matrices) by this method is improved for larger motions.

This sort of data combination can be applied to any situation where we know a pair of datasets are related by the same transformation as another pair. Thus we can apply it to solving the degeneracy inherent in calculating 3D projective transformations for planar structure.

If we know that the projective structure from one pair of stereo views is related by the same $T$ as another pair, we can combine the structures as if there were two planes in projective space undergoing transformation $T$ rather than one. This can be achieved with our stereo head because its odometry enables it to make precise relative motions about its axes, e.g. if required it can rotate repeatedly by the same angle about its pan axis. Note that the absolute angular value is not required, just repeatability; thus the method can be used by any head which can make repeatable motions, such as those driven with uncalibrated stepper motors.

5. Implementation

Here we sum up the major steps of the planar calibration routine.

1. Make a minimum of two rotations of the head about the pan axis by the same angle, followed by at least two rotations about the elevation axis. Taking images before and after each motion will result in a minimum of five stereo image pairs.

2. Detect point features in each image, and by matching them between left and right images calculate the planar homography $H_i$ relating these images for each view $i$. Also calculate the homographies $H_{ij}^l$, $H_{ij}^r$ that relate points in the left or right images of two stereo pairs.

3. By combining all the left image data, and all the right image data, calculate the epipolar geometry and hence projective structure.

4. For the pan and elevation motion in turn, combine the structure from all views related by the same motion as described in §4. Calculate the two $T$ matrices representing the pan and elevation motions (using, for instance, the algorithm described in [5]).

5. Mutually decompose the two $T$ matrices to give calibration and head geometry [5].

Calculating the epipolar geometry in step 3 can be done by explicit use of the combined feature data. However, methods exist for calculating structure for piecewise planar scenes (as in the combined view) using the planar homographies $H_i$ [7, 8]. This is more accurate since planarity is then being enforced.

Once we have image relationships between all images in a set of four (left and right, before and after a motion), we no longer need real image data. We can choose any set of points in one image and find their matches in the others using these homographies. In our implementation we simply choose the four corners of one of the images. Not only does this enforce planarity of backprojected structure (because left-right point matches are related exactly by a homography), giving improved accuracy, but it speeds up processing because there are a near minimal number of points.
the problem of self-calibration, it enables us to use general scene calibration algorithms in spite of the scene degeneracy. Consequently there is an advantage over previously reported planar calibration routines in that the scene is unknown, there are no initialisation problems, and no constraints must be made on the calibration. In simulation and real scene experiments we note some restrictions required to ensure good accuracy, but demonstrate the validity and simplicity of the system.

7. Summary and Conclusions

We have described a system for eliminating the degeneracy inherent in the projective geometry of a planar scene viewed with stereo cameras. It relies only on the ability of the stereo head to make repeatable motions about its axes, as most pan-tilt systems can. Applying the technique to the problem of self-calibration, it enables us to use general scene calibration algorithms in spite of the scene degeneracy. Consequently there is an advantage over previously reported planar calibration routines in that the scene is unknown, there are no initialisation problems, and no constraints must be made on the calibration. In simulation and real scene experiments we note some restrictions required to ensure good accuracy, but demonstrate the validity and simplicity of the system.

Acknowledgements

This work is supported by the UK’s Engineering and Physical Sciences Research Council through a studentship to JGKH, an Advanced Research Fellowship to IDR, and through grant GR/L58668.

References