MIT Media Laboratory, Perceptual Computing Technical Report #341
Submitted to the IEEE Symposium on Computer Vision

Camera self-calibration from one point correspondence

Ali Azarbayejani and **Alex Pentland**
Media Laboratory
Massachusetts Institute of Technology
Cambridge, MA 02139
U. S. A.

Abstract

The internal and relative orientations of one or more cameras can be obtained by tracking a single point from each of two or more cameras. Thus, for example, a stereo rig can be internally and relatively calibrated by walking around in front of it, or a single camera can be internally calibrated this way using a second camera for data collection.

The implications of this fact are important for bringing 3-D computer vision systems out of the laboratory and into practical use. Cumbersome calibration objects and all but the simplest image processing are not required. The accuracy of calibration that can be achieved has no theoretical limit since an arbitrarily large amount of data can be collected. Also, our calibration procedure addresses problems that are difficult or impossible for traditional calibration techniques, including wide-baseline stereo.

We present a solution technique for self-calibration from one tracked point and verify its viability on synthetic and real stereo rigs.

1 Self-calibration is crucial for practical vision systems

Most computer vision algorithms for 3-D measurement begin by assuming that camera parameters are known in advance. And why not? After all, if you have a camera that you are planning to use as a sensor for measurement, calibration is clearly something that is possible to do ahead of time to make the rest of your task that much easier.

But, as a practical matter, current calibration techniques are time-consuming processes that usually require special equipment, special calibration objects, and a great deal of care. As computer vision systems move out of the laboratory and into the hands of non-expert computer users, it becomes increasingly crucial that camera systems can self-calibrate, they can do so simply and accurately, and in the context of a variety of settings and tasks.

Fig. 1. It is not reasonable to expect traditional stereo correspondence methods to find a large number of matches in images like this, but it is reasonable to expect being able to reliably track a few points on a person with a considerable degree of uncertainty. Our calibration method applies to both sources of correspondence, allowing human movement itself to be a viable source of calibration data.

Self-calibration requires establishing a relatively large number of correspondences (usually points) and computing the geometry from these constraints. In stereo, correspondences for calibration are usually established by finding matches in the two images. In some circumstances, point and line matches are plentiful, but there are many instances when these traditional stereo techniques are not effective.

In the case of wide-baseline stereo, traditional image-based correspondence fails because there is not as much useful overlap and features are often too distorted to match reliably. In the extreme case of some 3-D monitoring systems, cameras encircle the subject, resulting in some cameras facing each other, in which case traditional stereo clearly does not apply.

Self-calibration is often thought of as not just a preprocess, but an ongoing process that happens "on-the-fly" as other vision tasks are taking place. This is especially relevant for mobile robots and active vision systems whose calibration can change during operation.

But some scenes simply do not allow a large number of correspondences at any given frame. Outdoor scenes, people, animals, and roads are examples of important vision applications in which reliable point and line matches are scarce.

For example, a typical image from a human interface is shown in Figure 1, along with a color classification result which is one step towards establishing a symbolic correspondence of human body parts. The rough segmentation of the human body suggests that tracking of a few major points of interest (like the center of the face, the left hand, etc...) are possible, with potentially a large degree of uncertainty. But certainly the dozens of precise point matches required for existing calibration techniques is unlikely, especially in cases where the view does not contain favorable background features as in this particular image.

Thus far we have argued that traditional stereo correspondence is not useful in certain situations such as self-calibration of wide-baseline systems, and self-calibration of some active or mobile systems. But even in cases where traditional stereo systems can do the job, sophisticated image processing is usually required, and is generally not portable to significantly different environments. Thus, the following insight on correspondence for self-calibration is applicable to traditional stereo in addition to some cases that are not handled well by traditional means.

1.1 Calibration from one correspondence per frame

The insight that we offer is that, for any two or more fixed cameras, any correspondence, taken at any frame, is a measurement of both the relative orientation and the internal orientation of each camera (and also the 3-D location of that point). Since the relative and internal orientations are static, tracking a single moving point to 10 different locations is the same as if you had matched 10 different points in those locations across a single image pair. In other words, one match in each of ten image pairs is the same as ten matches in one image pair.

So why bother with difficult image processing problems? Tracking a flashlight from one corner of the room to another effortlessly yields a large number of point correspondences, which can often be better than surface points because of a wider possible depth range. Literally any point in open space that is visible from both cameras can be made into a correspondence.

Thus, there are several important advantages offered by our technique that we summarize here.

First, if calibration is the only goal, data collection is trivial. If you can track a small light in a dark room, there is no limit to the amount of data that can be collected. With NTSC video, nearly 2000 correspondences can be obtained in one minute and the only image processing required is thresholding.

Second, there is no limit to the number of 3-D points that can be the source of correspondence. In contrast, when calibration objects or existing static landmarks are used at a single time instance the range of depths in the image is typically small compared to the visible volume. Furthermore, only one depth at any location in the image can be used for data because of occlusion. By tracking a point, an arbitrarily dense and broad set of data can be obtained in the volume.

Third, point tracking works in situations where iconic stereo matching cannot. Wide-baseline stereo and encircling multiple-camera systems are good examples of practical problems that our technique addresses and traditional stereo does not.

Finally, in cases such as human tracking, or road following, where symbolic correspondence can yield a few reliable, yet imprecise, matches, our technique offers a means of dynamic self-calibration.

1.2 Some relevant previous work

In 1986, Tsai [8] gave a thorough review of the calibration field at the time. Since then, calibration research has gone in many directions, primarily solving restricted parts of the problem, with few treatments of self-calibration for cameras in arbitrary configurations.

A seminal work on self-calibration of stereo rigs appeared the same year, by Faugeras and Toscani [3]. The authors touched on the topic of temporal integration of stereo correspondences as they described an extended Kalman filter for recursively computing the self-calibration from point correspondences on a moving stereo rig.

Our work pushes the concept beyond the intent of this previous work by suggesting that not only is temporal integration of stereo matches a good supplement to traditional spatial integration, but that the two sources of data are in fact identical and one can be replaced by the other with no detriment to performance. The implication is that establishing a few (or one) good landmarks can be as valuable as establishing a large number of possibly unreliable matches.

It is worth noting that Faugeras's stereo method could trivially be adapted to our data and is therefore the same functionally as our estimator. However, The Faugeras work was performed using the epipolar constraint for computation of relative and interior orientation. The epipolar constraint represents one tradition of motion and stereo computation that is quite widespread, but epipolar constraints only solve part of the geometry problem; they use only half the data and leave out the structure parameters. Because of this, it is often reported, as in a later calibration work [2], that estimation based on epipolar constraints can be unreliable in the presence of significant measurement noise.

Thus, our computational approach, which is in the tradition of complete nonlinear forward-model inversion, as in [4; 6; 5; 1; 7], offers an alternative which may turn out to be more successful at obtaining an accurate metric solution in the presence of imprecise measurements.

2 Solution via forward model inversion

To some readers, it may be obvious that waving a flashlight around, driving down a road, or watching a person move can provide the required data for self-calibration, but for others, and to motivate our solution, we turn here to a brief discussion of related computer vision topics concerned with computing *geometry from correspondence*.

2.1 Motivation — geometry from correspondence

The forward model of correspondence between two images of a point involves three stages. First, a 2-D point in one image is back-projected to a 3-D point using the internal orientation of the camera and the depth of the point, i.e. the structure. Second, the 3-D point is transformed into the other camera frame using the relative orientation. And, finally, the 3-D point is projected into the new image, using the new 3-D location and the internal orientation of the new camera.

This forward model can be computed for a single camera that moves (structure from motion, SfM) or for two static cameras (stereo self-calibration). The forward model and the inversion process are identical for the two cases, but parameters are treated differently. In structure from motion, the structure and internal orientation are usually static, while relative orientation is dynamic. In self-calibration, where correspondences are taken at one time instance, all parameters are static. In our version, when points are being tracked, the internal and relative orientations are static, while the structure parameters are dynamic.

Underneath the differences in parameter behavior, however, the geometric models are the same and the measurements (correspondences) are the same. Thus, computationally, we might ultimately imagine a unified framework for solving both problems.

For now, it is sufficient to realize that the underlying geometry of self-calibration is the same as SfM. This is a well-known fact in computer vision, but few researchers take advantage of it to actually use the same mechanism for both problems. We demonstrate here, by using existing SfM code, that a computational mechanism designed to accurately solve the SfM problem is also naturally well-suited to accurately solving the self-calibration problem.

2.2 Forward model inversion using the iterated Kalman filter

We directly adapt the underlying models of structure from motion estimation that have been argued [1] to be particularly well-suited to computational stability for the correspondence-based estimation problem. We use slightly different priors and we iterate, since the relative orientation can be large, but otherwise use the SfM estimator "off-the-shelf".

The SfM estimator is a Kalman filter based algorithm that operates by inverting the forward model, which has three major components. We describe them using stereo terminology (left and right cameras) although there is obviously a motion interpretation as well.

The three components of the underlying forward model are as follows: Back projection from the left image point to a 3-D point in the left camera frame uses the structure (depth) and internal orientation of the left camera, and is given by

$$\begin{pmatrix} X_L \\ Y_L \\ Z_L \end{pmatrix} = \begin{pmatrix} u_L \\ v_L \\ 0 \end{pmatrix} + \alpha_L \begin{pmatrix} u_L\beta_L \\ v_L\beta_L \\ 1 \end{pmatrix} \tag{1}$$

where (u_L, v_L) is the left image point, (X_L, Y_L, Z_L) is the 3-D point in the left camera frame, α_L is the depth, and β_L is the inverse focal length.

Frame transformation to the (distorted) right camera frame uses the relative orientation and right camera internal orientation and is given by

$$\begin{pmatrix} X_R \\ Y_R \\ Z_R\beta_R \end{pmatrix} = \begin{pmatrix} t_X \\ t_Y \\ t_Z\beta_R \end{pmatrix} + \mathbf{R}_b \begin{pmatrix} X_L \\ Y_L \\ Z_L \end{pmatrix} \tag{2}$$

where $(X_R, Y_R, Z_R\beta_R)$ is the distorted 3-D point in the right camera frame, $(t_X, t_Y, t_Z\beta_R)$ is the distorted translation, and

$$\mathbf{R}_b \doteq \begin{pmatrix} 1 & & \\ & 1 & \\ & & \beta_R \end{pmatrix} \begin{pmatrix} r_{11} & r_{12} & r_{13} \\ r_{21} & r_{22} & r_{33} \\ r_{31} & r_{32} & r_{33} \end{pmatrix}$$

is the distorted rotation matrix.

Finally, forward projection into the right image is given by

$$\begin{pmatrix} u_R \\ v_R \end{pmatrix} = \begin{pmatrix} X_R \\ Y_R \end{pmatrix} \frac{1}{1 + Z_R\beta_R}. \tag{3}$$

The modeling used here is not too far from a traditional treatment. Certain parameter choices are made for numerical reasons as discussed in the references.

The internal camera model is a pinhole model, parameterized with inverse focal length only. Image center, scaling, and radial distortion are omitted in this implementation. Any attempt at high precision calibration would have to eventually include these higher order parameters, but it is probably not possible to estimate these until until a rough calibration is first established as a starting point. The issue of high precision internal calibration using noisy, but plentiful, correspondence data is an interesting possibility for future research. Most existing internal calibration techniques use high precision measurements.

The state vector of the Kalman filter consists of the relative orientation, the internal orientations, and the structure parameters.

$$\mathbf{x} = (t_X, t_Y, t_Z\beta_R, \mathbf{R}_b, \beta_R, \beta_L, \alpha_{L1}, \cdots, \alpha_{LN}) \tag{4}$$

where N is the number of points collected. (The rotation matrix is not actually used in the state vector, but rather angular interframe rotations are computed; see [1] for details.)

The independent parameters are the left image coordinages (u_L, v_L) and the measurements are the right image coordinates (u_R, v_R).

The way we performed our experiments was to collect a set of data and compute one time step of the SfM filter, iterating until convergence. There are alternatives, such as recursively incorporating each measurement on-the-fly when it is obtained. This is another topic for future work.

3 Experiments

To demonstrate the estimator performance, we have conducted simulated experiments on wide baseline stereo configurations, of which we show here a particularly difficult case. We also demonstrate the real-world use of the estimator by calibrating a wide-baseline stereo rig using nothing more than a flashlight and binary image processing on thresholded images.

3.1 Simulation performance

Figure 2 shows the results of a particularly difficult wide-baseline relative orientation calibration problem in which two cameras observe a space from orthogonal directions. The two cameras are located on a circle, rotated 90° from each other with their optical axes intersecting orthogonally at the center of the circle.

In this experiment, the focal length was known and the relative orientation (and structure of the calibration points) were computed.

The correct values for the translation are $(-2, 0, 2)$ and for the rotation quaternion are $(.7071, 0, .7071, 0)$. The structure parameters are clustered at five depths, $(1, 1.2929, 2, 2.7071, 3)$.

The data was generated synthetically and no noise was added. It is a difficult problem because of the large rotation. The iterated extended Kalman filter converges in about 20 frames.

3.2 Extremely practical stereo calibration

We demonstrate the practicality of our calibration method by calibrating a wide-baseline stereo system. We use a data set of 100 points obtained by waving the "magic wand", i.e. a flashlight, in the visible volume to compute the calibration parameters and then testing the result by triangulating two additional sets of data representing a recognizable object.

In Figure 3, our stereo system is shown at the top. It consists of two off-the-shelf color video CCD cameras mounted overlooking a person's work space. No particular care is taken in mounting the cameras, other than that they both adequately cover the human subject in their fields of view.

The pair of pictures shows the scene, including a squarish video monitor on a chair and the flashlight which was used for taking calibration and test data. The next two pictures show the feature traces of the flashlight in both cameras. These are the input to the iterated Kalman filter which computes the calibration parameters shown in the graphs.

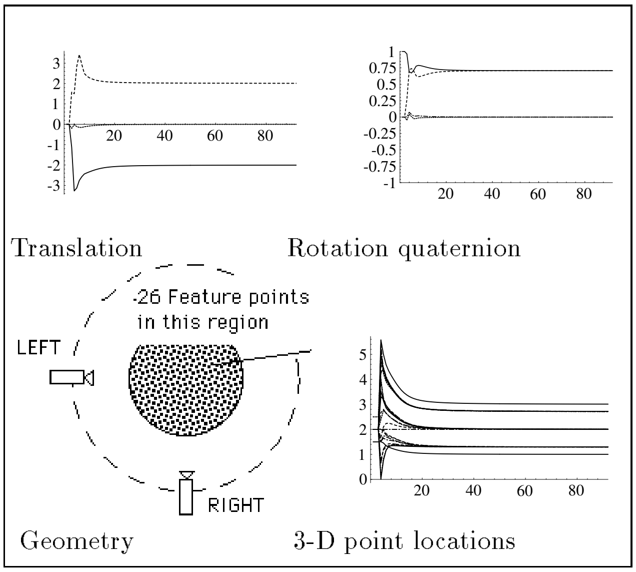

Fig. 2. Calibration of a synthetic wide-angle stereo rig using 26 correspondences.

The calibration is verified by taking a new set of correspondences with the flashlight and computing the 3-D points. We ran the light along the edges of the top and front of the box (a video monitor), which is roughly square. The bottom picture is a stereo pair of the recovered 3-D data, which can be visually fused by the reader to perceive the 3-D points.

There was a quite a bit of error in the movement of the flashlight relative to the edges of the box; the data should be interpreted as a very rough outline of the top and front of the box, not to be judged locally, but rather globally.

We have no ground truth with which to compare the box data, but visual inspection verifies that the recovered points representing the squarish top and front both look squarish and the two surfaces are orthogonal to each other. Scale is actually not recoverable in stereo self-calibration, so absolute distances have no meaning anyway until they are rectified with actual lengths. In this case, the object is 35 cm wide and 37 cm deep.

The accuracy observed in the reconstruction is sufficient accuracy for many interesting applications, and certainly worth the cost. Data collection took 5 seconds and computation of 3-D parameters took another 15 seconds, and no image processing beyond simple thresholding was required.

4 Discussion

We have discussed the issue of self-calibration from a single tracked point. We used an off-the-shelf correspondence-based estimator to calibrate a wide-baseline stereo system using nothing more than a flash-

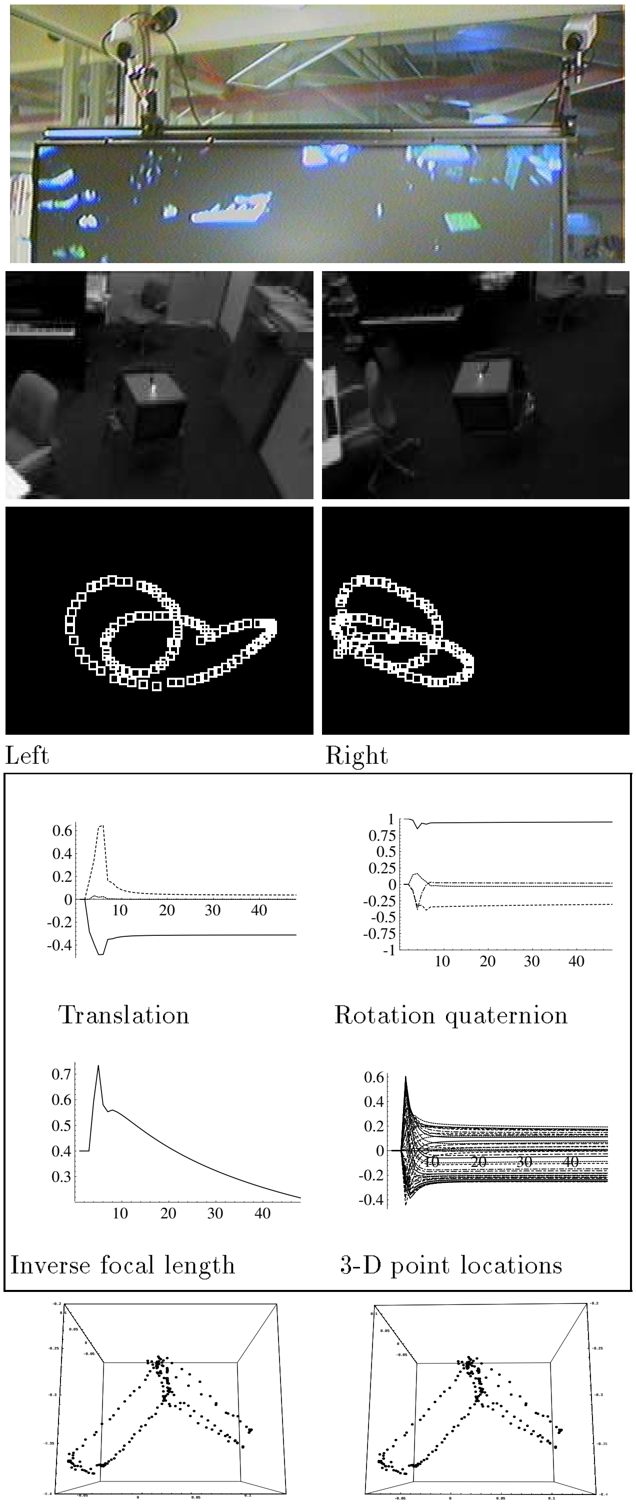

Fig. 3. Calibration of a real stereo rig by tracking a single point. The views from the left and right camera show the flashlight used to obtain point correspondences and the monitor which was scanned in using the calibration data. The graphs show the iterations of the IEKF. At bottom is a stereo pair of the 3-D points obtained by moving the flashlight along the edges of the CRT monitor.

light and threshold-based image processing.

Our results demonstrate the value of our technique for practical vision systems. A multiple-camera system can be plunked down virtually anywhere it is needed and calibrated in seconds or minutes with a small set of standard software.

4.1 Future Work

The ideas presented in this paper can be extended in a number useful ways.

First, the technique is not particular to stereo. The calibration technique can be naturally extended to multiple cameras. The technique can also be extended to the internal orientation of a single camera if a second camera can be used for data collection.

Second, the technique may be extended to more sophisticated internal camera models, as used in high-precision internal calibration. Existing techniques usually rely on high precision measurements. Since the valuable characteristic of our technique is the availability of an abundance of data, not necessarily the precision of the data, the challenge here is to extract the useful information from large quantities of imprecise data.

Finally, although we collected our data temporally we did not process it recursively. An obvious extension, then, is to process measurements recursively. This is a necessary step for applications requiring self-calibration on-the-fly.

Acknowledgments

The authors would like to acknowledge the help of Lee Campbell, Tom Minka, and Steve Mann.

References

[1] Ali Azarbayejani and Alex Pentland. Recursive estimation of motion, structure, and focal length. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 17(6), June 1995. To Appear. (Also, Media Lab TR 243, http://www-white.media.mit.edu/vismod).

[2] O. D. Faugeras, Q.-T. Luong, and S. J. Maybank. Camera self-calibration: Theory and experiments. In G. Sandini, editor, *Computer Vision, ECCV '92*, pages 321–334, Berlin, May 1992. DIST, University of Genova, Springer-Verlag. (Santa Margherita Ligure, Italy).

[3] O. D. Faugeras and G. Toscani. The calibration problem for stereo. In *1986 IEEE Conference on Computer Vision and Pattern Recognition*, pages 15–20, Los Alamitos, CA, June 1986. IEEE Computer Society, IEEE Computer Society Press. (Miami Beach).

[4] Berthold Klaus Paul Horn. Relative orientation. *International Journal of Computer Vision*, 4(1):59–78, January 1990.

[5] Larry Matthies, Takeo Kanade, and Richard Szeliski. Kalman filter based algorithms for estimating depth

from image sequences. *International Journal of Computer Vision*, 3(3):209–236, 1989.

[6] Minas Spetsakis and Yiannis Aloimonos. A multi-frame approach to visual motion perception. *International Journal of Computer Vision*, 6(3):245–255, August 1991.

[7] Richard Szeliski and Sing Bing Kang. Recovering 3d shape and motion from image streams using non-linear least squares. In *1993 IEEE Conference on Computer Vision and Pattern Recognition*, pages 752–753, Los Alamitos, CA, June 1993. IEEE Computer Society, IEEE Computer Society Press. (New York).

[8] Roger Y. Tsai. An efficient and accurate camera calibration technique for 3d machine vision. In *1986 IEEE Conference on Computer Vision and Pattern Recognition*, pages 364–374, Los Alamitos, CA, June 1986. IEEE Computer Society, IEEE Computer Society Press. (Miami Beach).