Learning priors for calibrating families of stereo cameras

Andrew W. Fitzgibbon
Microsoft Research
awf@microsoft.com

Duncan P. Robertson
Microsoft Research
duncan.robertson@redimension.co.uk

Antonio Criminisi
Microsoft Research
antcrim@microsoft.com

Srikumar Ramalingam
Oxford Brookes University
srikumar.ramalingam@brookes.ac.uk

Andrew Blake
Microsoft Research
ablake@microsoft.com

Abstract

Online camera recalibration is necessary for long-term deployment of computer vision systems. Existing algorithms assume that the source of recalibration information is a set of features in a general 3D scene; and that enough features are observed that the calibration problem is well-constrained. However, these assumptions are frequently invalid outside the laboratory. Real-world scenes often lack texture, contain repeated texture, or are mostly planar, making calibration difficult or impossible.

*In this paper we consider the calibration of families of stereo cameras, where each camera is assumed to have parameters drawn from a common but unknown prior distribution. We show how estimation of this prior using a small-number of offline-calibrated cameras (*e.g. *from the same production line) allows online calibration of additional cameras using a small number of point correspondences; and that using the estimated prior significantly increases the accuracy and robustness of stereo camera calibration.*

1. Introduction

Computer vision systems are increasingly deployed in mass-market and consumer applications, where high performance is desired in a wide range of uncontrolled and unprepared environments. In such contexts, the problem of camera calibration is especially challenging. Here, we consider stereo webcams [8, 10, 12], where the calibration of the epipolar geometry is an important prerequisite for many useful applications. One possibility is to calibrate cameras in the factory. However, this is expensive, particularly for high production volumes, and might be insufficient if the stereo geometry changes over time, *e.g.* due to thermal or mechanical phenomena. For these reasons, the preferred approach is usually to calibrate the camera *online*, *i.e.* using correspondence data obtained directly from the scene—either when the user installs the camera or at regular intervals thereafter.

A significant limitation of conventional calibration algorithms is the assumption of rich 3D scene structure, which may not be available in a significant proportion of real-world scenes, particularly in homes and offices. Even if the user is provided with a calibration target, obtaining a suitably rich set of correspondences throughout a camera's working volume can prove surprisingly difficult. In this paper we argue that better calibration results can be obtained more easily by employing strong priors in a Bayesian framework (figure 1). In such a framework, the prior calibration is combined with the data likelihood (possibly computed from only a small number of degenerate scene correspondences) to obtain an accurate posterior camera calibration. This approach will be shown to be more robust (*i.e.* it works in structure-poor scenes) and more accurate (*i.e.* it produces smaller errors) than calibration by conventional means.

This paper's main concern is with how to estimate priors for camera calibration parameters given a small number of stereo cameras from the same production line and their associated offline calibrations. It is well known *how* to incorporate priors into camera calibration [17, 21], but the problem of obtaining the prior in the first place has seen little or no attention.

The contributions of this paper are as follows: first, introducing the problem of finding good priors for stereo rig calibration; second, estimation of the covariance matrix of a small number of uncertainly-known points; third, the experimental comparison of various priors for stereo camera calibration.

1.1. Background

Two strands of literature are relevant to our investigation: first is Bayesian tracking of camera motion [1, 3, 5, 9, 14]; second is the work on degeneracy in structure and motion recovery [13, 20].

Bayesian tracking of camera calibration and motion has largely been addressed in the context of mobile robotics, where a camera moves through a 3D scene, maintaining an estimate of the probability distribution over its camera

Figure 1. **Stereo camera calibration results in typical structure-poor scenes.** Four image pairs were obtained by a single stereo camera (only right-eye views are shown) and stereo correspondences were obtained by robust feature matching. Without a prior, standard camera calibration techniques fail and the recovered epipolar geometry (shown using epipolar lines) is inaccurate (top row). By using a prior learned by the method described in this paper, accurate (and consistent) calibration is obtained (bottom row).

parameters and position. Research has focussed on efficient maintenance of the distributions, e.g. by Kalman filter [1, 3, 5, 9], or particle filter [14], rather than on the construction of the prior, which at the initial time step is either generic [5, 9] or obtained by a batch process [3]. The former case is close to our problem, because the estimate of calibration is only weakly constrained for the initial steps. However, the essential assumption of Bayesian tracking is that the scene geometry is rich enough that the estimate will be constrained after a relatively small number of time steps. In home and office scenes, however, we cannot assume that this will ever be the case.

More directly related to our study of structure-deficient calibration is the work on coping with degeneracy in structure and motion recovery. Torr *et al*. [20] took earlier theoretical work on degeneracy and showed how detection of structure-poor sections of an image sequence allowed point tracking to be maintained through degeneracy, specifically low-relief scenes. Pollefeys *et al*. [13] went further, showing how 3D structure could be recovered despite planar subsequences, but only if the camera eventually views some rich 3D structure.

Many publications specifically deal with "online calibration", e.g. [4, 16]. In all the works of which we are aware, the online calibration requires the same rich 3D structure as the above-cited works, or requires additional information such as controlled motion, vanishing points, or other manually supplied information. Thus no existing work deals with the case of sequences which are structure-poor throughout. Indeed we have been unable to find any work which deals with building priors for camera calibration. The existing mentions [1, 9] all assume generic priors which are not adequate for this problem, as the experiments in this paper show.

1.2. Notation

We begin by defining notation for the problem. A camera system[1] is characterized by a vector of parameters $\boldsymbol{\theta}$. The parameters might be, for example, the seven parameters describing the focal length and 3D position of a security camera. In this work, $\boldsymbol{\theta}$ will contain the intrinsic and extrinsic parameters of a stereo rig. The choice of parameterization will be discussed in Section 3.

The cameras are drawn from a prior distribution (a camera *family*) that will be approximated by a Gaussian with parameters $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$, giving a prior pdf

$$p(\boldsymbol{\theta}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \mathcal{N}(\boldsymbol{\theta}|\boldsymbol{\mu}, \boldsymbol{\Sigma}). \tag{1}$$

Before any estimation, we have *a-priori* estimates of the values of $(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, for example from the manufacturing blueprints, which we shall denote by $(\boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0)$. As the goal of the paper is to compute an estimate of $\boldsymbol{\mu}, \boldsymbol{\Sigma}$, we also introduce a hyper-prior on these parameters. The hyper-prior on $\boldsymbol{\mu}$ is taken to be a broad Gaussian with mean $\boldsymbol{\mu}_0$ and precision γ, and the hyper-prior on $\boldsymbol{\Sigma}$ to be an inverse Wishart distribution [2] with mean $\boldsymbol{\Sigma}_0$. We shall work in a scaled space, so that $\boldsymbol{\Sigma}_0$ is the identity, meaning that the hyper-prior simplifies to

$$p(\boldsymbol{\Sigma}|\boldsymbol{\Sigma}_0)\Big|_{\boldsymbol{\Sigma}_0=\mathbf{I}} \propto |\boldsymbol{\Sigma}|^{-\frac{1}{2}(\nu-d-1)} \exp\left(-\frac{1}{2}\mathrm{Tr}(\boldsymbol{\Sigma}^{-1})\right) \tag{2}$$

[1] Note that we use the word *camera* in the generalized sense of [6, 18], to refer to any ray-gathering device. Throughout this paper, "camera" and "stereo rig" may be considered to be interchangeable terms.

where d is the dimensionality of $\boldsymbol{\theta}$ and ν is the degrees of freedom of the Wishart distribution.

Image measurements are a set of observations $\mathcal{Z} = \{\mathbf{z}_i\}_{i=1}^N$, where the $\mathbf{z}_i$ are point correspondences between two images, represented as 4-vectors $\mathbf{z} = (u, v, u', v')$. Associated with each observation is a latent variable $\mathbf{x}_i$, which here will be the 3D point which gave rise to the correspondence. Finally there is a projection function $f(\boldsymbol{\theta}, \mathbf{x})$ which generates noiseless observations, to which are added noise $\boldsymbol{\eta}$ drawn from a Gaussian noise distribution with density function $p(\boldsymbol{\eta}) \propto \exp(-\frac{1}{2}\|\boldsymbol{\eta}\|^2/\sigma^2)$, where image coordinates are assumed scaled so that $\sigma = 1$. Thus

$$\mathbf{z} = f(\boldsymbol{\theta}, \mathbf{x}) + \boldsymbol{\eta} \tag{3}$$

is our observation model.

In both offline and online calibration, we will optimize image measurements under a Gaussian prior in order to approximate the distribution $p(\boldsymbol{\theta}|\mathcal{Z})$. Expanding $p(\boldsymbol{\theta}|\mathcal{Z})$ to the product of likelihood and prior,

$$p(\boldsymbol{\theta}|\mathcal{Z}, \boldsymbol{\mu}, \boldsymbol{\Sigma}) \propto p(\mathcal{Z}|\boldsymbol{\theta}) p(\boldsymbol{\theta}|\boldsymbol{\mu}, \boldsymbol{\Sigma}). \tag{4}$$

The likelihood $p(\mathcal{Z}|\boldsymbol{\theta})$ is obtained by optimizing over the latent variables $\mathbf{x}$ in (3):

$$p(\mathcal{Z}|\boldsymbol{\theta}) = \exp\left\{-\frac{1}{2}\sum_{i=1}^{N} \min_{\mathbf{x}} \|\mathbf{z}_i - f(\boldsymbol{\theta}, \mathbf{x})\|^2\right\}, \tag{5}$$

and the prior is the Gaussian $p(\boldsymbol{\theta}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) =$

$$\det(2\pi\boldsymbol{\Sigma})^{-\frac{1}{2}} \exp\left\{-\frac{1}{2}(\boldsymbol{\theta} - \boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{\theta} - \boldsymbol{\mu})\right\}. \tag{6}$$

The negative log posterior –$\log p(\boldsymbol{\theta}|\mathcal{Z})$ then becomes

$$\epsilon(\boldsymbol{\theta}; \mathcal{Z}, \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \sum_{i=1}^{N} \min_{\mathbf{x}} \|\mathbf{z}_i - f(\boldsymbol{\theta}, \mathbf{x})\|^2 + \\ + (\boldsymbol{\theta} - \boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{\theta} - \boldsymbol{\mu}) \tag{7}$$

where immaterial constants have been omitted. Minimizing this negative log posterior is a standard problem, solved by bundle adjustment algorithms [7, 21]. For online calibration, we will be interested in the minimum, under a suitable prior.

For offline calibration, in which the prior is computed, we will also require a Laplace-like approximation [2, p214ff] to the likelihood (5), *as a function of* $\boldsymbol{\theta}$, in which

$$p(\mathcal{Z}|\boldsymbol{\theta}) \approx \zeta \mathcal{N}(\boldsymbol{\theta}|\boldsymbol{\mu}_{\text{Laplace}}, \boldsymbol{\Sigma}_{\text{Laplace}}) \tag{8}$$

where ζ removes the scale factor which normalizes for $\mathcal{Z}$ rather than $\boldsymbol{\theta}$. Then

$$\boldsymbol{\mu}_{\text{Laplace}} := \operatorname*{argmin}_{\boldsymbol{\theta}} \epsilon(\boldsymbol{\theta}; \mathcal{Z}, \boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0) \tag{9}$$

$$\boldsymbol{\Sigma}_{\text{Laplace}} := \mathbf{H}^{-1} \tag{10}$$

and $\mathbf{H}$ is the Hessian of $-\log p(\mathcal{Z}|\boldsymbol{\theta})$ evaluated at $\boldsymbol{\mu}_{\text{Laplace}}$, with ij^{th} entry $H_{ij} = \frac{\partial}{\partial\theta_j}\frac{\partial\epsilon}{\partial\theta_j}$.

2. Algorithm overview

With the mathematical preliminaries in place, we are now in a position to outline the calibration strategy proposed in this paper. Recall that the goal is to compute prior parameters $(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ in an offline stage so that online calibration using a small number of correspondences is accurate despite degeneracy in the scene.

2.1. Offline phase: estimating the prior

The offline process is the primary focus of this paper. In principle this is straightforward: we are given M cameras, and for each camera, indexed by m, we obtain a large, space-filling set of correspondences $\mathcal{Z}_m = \{\mathbf{z}_{mi}\}_{i=1}^M$. From each set of correspondences, compute $\boldsymbol{\theta}_m = \operatorname{argmax}_{\boldsymbol{\theta}} p(\boldsymbol{\theta}|\mathcal{Z}_m, \mu_0, \boldsymbol{\Sigma}_0)$ from (5). Then the estimate of the prior mean is the **sample mean** $\boldsymbol{\mu} = \frac{1}{M}\sum_m \boldsymbol{\theta}_m$, and the prior covariance is the **sample covariance** $\boldsymbol{\Sigma} = \frac{1}{M-1}\sum_m (\boldsymbol{\theta}_m - \boldsymbol{\mu})(\boldsymbol{\theta}_m - \boldsymbol{\mu})^\top$. In practice there are a number of factors that make this a challenging problem, rendering these estimates essentially useless.

The primary difficulty is that obtaining and calibrating enough cameras to compute a reliable covariance matrix using these standard estimators is a considerable effort. For a stereo rig with pinhole cameras, the minimum number of parameters needed to characterize the rig is seven, for the seven degrees of freedom of the fundamental matrix. Estimation of the 28 parameters of the prior covariance matrix $\boldsymbol{\Sigma}$ requires at least eight cameras, and in practice considerably more. Furthermore, parametrizing the geometry by a minimal parametrization will generally mean that the distribution of camera parameters (i.e. fundamental matrices) in $\mathbb{R}^7$ is highly nonlinear, and thus unlikely to be well modelled by a Gaussian. In this work we choose instead to parametrize the geometry in $\mathbb{R}^{12}$ as described in §3, leading to a distribution which is better approximated by a Gaussian, but now requires many tens of offline calibrations in order to estimate $\boldsymbol{\Sigma}$ using conventional techniques. Indeed the sample covariance matrix will often not even be of full rank.

In practice we are limited to being able to calibrate only a small number of cameras—in the experiments here we used ten. This is largely because accurate camera calibration is nontrivial, even in a laboratory setting. Svoboda [19], for example, describes how point correspondences are generated between the cameras by moving a target such as a light source, so as to densely sample the 3D workspace, and then the camera parameters are found by maximizing the above likelihood. In our experience, such calibrations are surprisingly difficult, and are rarely successful on the first

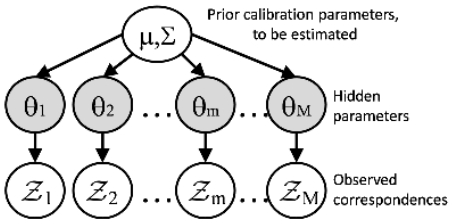

Figure 2. **Graphical model**. Illustrating the estimation of the prior from the offline calibrations. The offline calibration parameter vectors $\boldsymbol{\theta}_{1..M}$, one vector per camera, are hidden variables which are integrated out to give a direct estimate of the prior parameters from the calibration data.

capture, even for technically adept users. Our offline procedure (§3) is simpler, but still requires considerable physical movement of the calibration target to get a good calibration.

As a result of these constraints, it is desirable to use this calibration data as efficiently as possible in estimating the prior. One way is to consider only a restricted class of covariance matrices, for example zeroing the off-diagonal terms of the sample covariance matrix to obtain a **diagonal sample covariance** $\boldsymbol{\Sigma}_d = \text{diag}(\bar{\boldsymbol{\Sigma}})$. This avoids the problems of rank-deficiency, but the number of samples available may still be too small to obtain a reliable estimate (as our experiments show).

Bayesian estimate of the prior

Rather than estimate the prior from the small number of samples $\boldsymbol{\theta}_m$, we return to the initial correspondence sets $\mathcal{Z}_{1..M}$. By directly optimizing the likelihood of the offline calibration data with respect to the prior parameters, a more robust estimate of the prior can be obtained. The likelihood of all correspondences sets is illustrated in the graphical model of figure 2, and is as follows.

Each camera has a set of parameters $\boldsymbol{\theta}_m$, drawn from the prior with parameters $(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, i.e. from the distribution

$$p(\boldsymbol{\theta}_m | \boldsymbol{\mu}, \boldsymbol{\Sigma}) \tag{11}$$

Then each correspondence set $\mathcal{Z}_m$ is dependent on the camera parameters, yielding a total likelihood of the form

$$p(\mathcal{Z}_m | \boldsymbol{\theta}_m, \boldsymbol{\mu}, \boldsymbol{\Sigma}) = p(\mathcal{Z}_m | \boldsymbol{\theta}_m) p(\boldsymbol{\theta}_m | \boldsymbol{\mu}, \boldsymbol{\Sigma}) \tag{12}$$

In order to estimate $(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, we first marginalize over $\boldsymbol{\theta}_m$, yielding

$$p(\mathcal{Z}_m | \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \int p(\mathcal{Z}_m | \boldsymbol{\theta}_m) p(\boldsymbol{\theta}_m | \boldsymbol{\mu}, \boldsymbol{\Sigma}) \mathrm{d}\boldsymbol{\theta}_m, \tag{13}$$

Given: M cameras, hyper-prior parameters $\boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0$.
Compute: Prior parameters $(\boldsymbol{\mu}, \boldsymbol{\Sigma})$.
Procedure:
A. For each camera $m = 1..M$

1. Identify point correspondences $\mathcal{Z}_m$.
2. Optimize (7) with priors $\boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0$ using bundle adjustment, giving $\boldsymbol{\mu}_m$.
3. Compute $\boldsymbol{\Sigma}_m$ from the bundle Hessian (10).

B. Compute $(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ by minimizing (17).

Figure 3. **Computing the prior**. Summary of the algorithm for offline computation of the prior calibration $(\boldsymbol{\mu}, \boldsymbol{\Sigma})$.

and then optimize for $(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, under the prior (2):

$$(\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \underset{(\boldsymbol{\mu}, \boldsymbol{\Sigma})}{\text{argmax}} \prod_m p(\mathcal{Z}_m | \boldsymbol{\mu}, \boldsymbol{\Sigma}) p(\boldsymbol{\mu} | \boldsymbol{\mu}_0) p(\boldsymbol{\Sigma} | \boldsymbol{\Sigma}_0). \tag{14}$$

Given a Gaussian approximation to $p(\mathcal{Z}_m | \boldsymbol{\theta}_m)$ of the form $\zeta \mathcal{N}(\boldsymbol{\theta}_m | \boldsymbol{\mu}_m, \boldsymbol{\Sigma}_m)$, the integral (13) has the form of a product of Gaussians

$$p(\mathcal{Z}_m | \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \int \zeta \mathcal{N}(\boldsymbol{\theta}_m | \boldsymbol{\mu}_m, \boldsymbol{\Sigma}_m) \mathcal{N}(\boldsymbol{\theta}_m | \boldsymbol{\mu}, \boldsymbol{\Sigma}) \mathrm{d}\boldsymbol{\theta}_m \tag{15}$$

$$= \zeta \mathcal{N}(0 | \boldsymbol{\mu}_m - \boldsymbol{\mu}, \boldsymbol{\Sigma}_m + \boldsymbol{\Sigma}), \tag{16}$$

a function of $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$. The maximization (14) becomes (after taking logs and discarding constants) a minimization of the negative log posterior $L(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ given by

$$\begin{aligned} L(\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \sum_{m=1}^{M} & \log \det(\boldsymbol{\Sigma}_m + \boldsymbol{\Sigma}) + \\ & + (\boldsymbol{\theta}_m - \boldsymbol{\mu})^\top (\boldsymbol{\Sigma}_m + \boldsymbol{\Sigma})^{-1} (\boldsymbol{\theta}_m - \boldsymbol{\mu}) + \\ & + \gamma \|\boldsymbol{\mu} - \boldsymbol{\mu}_0\|^2 - 2 \log p(\boldsymbol{\Sigma} | \boldsymbol{\Sigma}_0), \end{aligned} \tag{17}$$

where $p(\boldsymbol{\Sigma} | \boldsymbol{\Sigma}_0)$ is the Wishart prior (2). Minimization of this function is discussed in §3. The offline algorithm is summarized in figure 3.

Having computed an offline prior, online calibration may be performed using a far smaller set of correspondences (including coplanar correspondences). During online calibration, camera parameters are obtained by minimizing (7) using the offline-computed prior.

3. Implementation

The previous sections have given a high-level overview of the procedure, the details of which will be filled out in

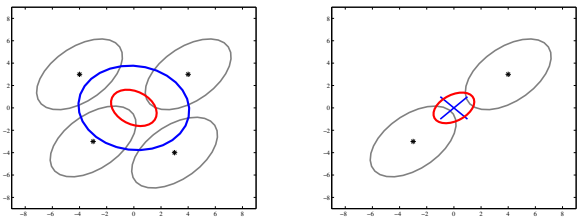

Figure 4. **Fitting a Gaussian to Gaussian samples**. (Left) Four samples and (right) two samples, represented by their modes $\boldsymbol{\mu}_m$ and associated covariances $\boldsymbol{\Sigma}_m$ are used to estimate a Gaussian in 2D. The sample covariance (blue) of the modes is overestimated if the number of samples is sufficent to estimate it, and cannot be estimated for small numbers of samples (blue $\times$ on the right). The Bayesian estimate derived in this paper (red) is robust in both cases.

this section. Important issues are the choice of parametrization, acquiring correspondences, initialization of the minimizer, and the setting of the hyper-prior covariance $\boldsymbol{\Sigma}_0$.

Parameterization. Our primary goal in choosing a parameterization for the stereo rig is to select a representation wherein the prior over camera parameters may be well modelled by a Gaussian. Rather than choosing a minimal parameterization (e.g. seven parameters for the fundamental matrix), we opt for an overparameterization, which is more likely to be smooth, and to allow the optimizer and prior to deal with gauge freedom. In the following experiments, we consider both Euclidean and projective parameterizations of the stereo camera geometry.

The **Euclidean parameterization** uses the following intrinsic parameters for each camera: focal length α, principal point (p_x, p_y). Optionally, one or two radial distortion coefficients can be included—for the cameras we tested, doing so makes no significant difference to our results. We assume the aspect ratio is known, and that the pixel skew is zero. The motion between the two cameras in a stereo rig is given by 3 rotational and 2 translational parameters. The rotation is represented by a three-parameter Rodrigues vector $\boldsymbol{\omega}$, with which is typically associated a base rotation matrix $\mathbf{R}_0$ as described in [21]. Let function $\mathrm{rod}(\boldsymbol{\omega})$ denote the 3×3 matrix representation of the rotation defined by $\boldsymbol{\omega}$. Translation is modelled as a three-vector $\mathbf{t}$. Three-dimensional points $\mathbf{x}$ are parameterized by homogeneous 4-vectors. Thus, the 12-element parameter vector is

$$\boldsymbol{\theta} = \left[\alpha, p_x, p_y, \alpha', p'_x, p'_y, \omega_1, \omega_2, \omega_3, t_x, t_y, t_z\right]^\top \quad (18)$$

where primed quantities are in the second camera. The observation function f may now be defined. This function takes parameters $\boldsymbol{\theta}$ and a 3D point $\mathbf{x}$, and generates a single

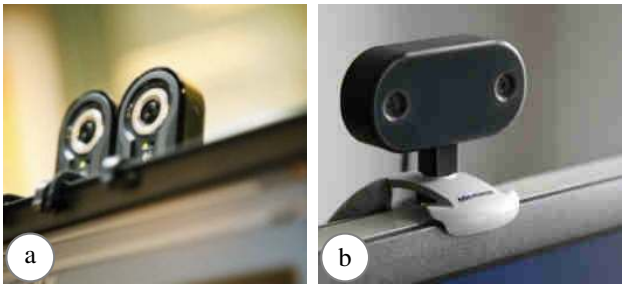

Figure 5. Stereo webcams. (a) "gluecam": one of 10 stereo webcams made by glueing together a pair of mono webcams. (b) "Hydra": a commercially available stereo webcam [10].

correspondence (x, y, x', y'), which is defined by

$$f(\boldsymbol{\theta}, \mathbf{x}) = \begin{bmatrix} \pi\left(\begin{bmatrix}\alpha & 0 & p_x \\ 0 & \alpha & p_y \\ 0 & 0 & 1\end{bmatrix}\left[\,\mathbf{I}\,|0\,\right]\,\mathbf{x}\right) \\ \pi\left(\begin{bmatrix}\alpha' & 0 & p'_x \\ 0 & \alpha' & p'_y \\ 0 & 0 & 1\end{bmatrix}\left[\mathrm{rod}(\boldsymbol{\omega})\mathbf{R}_0\,|\mathbf{t}\,\right]\,\mathbf{x}\right) \end{bmatrix} \quad (19)$$

where $\pi([x, y, z]) := [x/z, y/z]$.

The **projective parameterization** is that used for the "Gold Standard" calibration algorithm described in [7]. Here, the first and second cameras are represented by 3×4 projection matrices $\mathbf{P}$ and $\mathbf{P}'$ where $\mathbf{P} = [\,\mathbf{I} \mid 0\,]$ is fixed and $\mathbf{P}'$ is allowed to vary. The parameter vector $\boldsymbol{\theta}$ comprises simply the 12 elements of $\mathbf{P}'$, and the observation function $f = [\pi(\mathbf{P}\mathbf{x}), \pi(\mathbf{P}'\mathbf{x})]$.

Acquiring correspondences. In the offline calibration of the m^{th} stereo camera, we obtain dense correspondences by moving a large textured planar target relative to the camera. Correspondences are obtained for each target position by detecting Harris corners, matching local 7×7 pixel image patches by normalized cross correlation, and robustly fitting a plane using RANSAC [7]. This gives reliable correspondences for each target position.

Hyper-prior parameters. The "blueprint" or ambient prior is defined by considering the geometry of the stereo camera, typically by inspection of datasheets and estimation of manufacturing tolerances. For example, the ambient prior for the first of the cameras used in our experiments, described below, was a diagonal Gaussian with mean $\boldsymbol{\mu}_0$ = [960,320,240,960,320,240,0,0,0,-1,0,0] and covariance matrix $\boldsymbol{\Sigma}_0$ =diag(400, 100, 100, 400, 100, 100, 0.001, 0.001, 0.001, 0.001, 0.001, 0.002). These quantities are expressed in pixel coordinates, but calculations are scaled so as to work in coordinates such that $\boldsymbol{\mu}_0 = 0$ and $\boldsymbol{\Sigma}_0$ is the identity. The Wishart parameter ν is set in all cases to 3000, being approximately the number of data points used to compute the offline estimates.

Minimization. The negative log posterior (17) is not convex with respect to $\boldsymbol{\Sigma}$, and its minimization by gradient-based methods therefore requires that an initial estimate be provided. For our experiments, we searched not over all positive definite matrices, but over a one-parameter family which is simply a scalar t times the diagonal of the sample covariance matrix $\boldsymbol{\Sigma}(t) = t\boldsymbol{\Sigma}_d$. We also experimented with another one-parameter family which is a regularized sample covariance matrix $\boldsymbol{\Sigma}(t) = (1-t)\bar{\boldsymbol{\Sigma}} + t\boldsymbol{\Sigma}_0$.

4. Results

We tested the calibration algorithm described in this paper using two different kinds of stereo camera. Firstly, we constructed ten stereo "gluecams" by taking low-cost webcams, all of the same make and model, and glueing them together (see figure 5a). The webcams were roughly cuboidal in shape so the glueing process was reasonably repeatable. Additionally, we obtained 10 commercially available stereo webcams [10] (see figure 5b). These "Hydra" cameras are made with two imaging arrays mounted on a single printed circuit board, which is in fixed position relative to the two lenses mounted into the aluminium case.

For each stereo camera we collected 2D point correspondences by placing a large, textured, planar target in four positions (facing the camera and approximately 10, 20, 30, and 40 baselines away). Concatenating correspondences over these plane positions yielded about 3000 in total, distributed throughout the working volume of each camera. Then offline calibration was performed using the method described in section 3.

To report on calibration results requires an **error metric** that can compare a given calibration $\boldsymbol{\theta}_{\text{online}}$ to the ground truth. In what follows, we evaluate the calibration of stereo cameras for which we have several thousand left-out correspondences as well as "ground truth" camera parameters $\boldsymbol{\theta}_{\text{offline}}$. One metric that might be considered is to compare the parameter vectors, but this implies that one has a reasonable model of the expected errors on these vectors. We prefer a metric that relates to the intended application of the stereo cameras, which is dense 3D reconstruction. In this context, performance is strongly correlated with the accuracy of the rectification, and hence with the distance of points to their corresponding epipolar lines. Thus in each case, we quote the "rectification error", denoted **RFE**, defined as the root mean square of the symmetric transfer distance defined by Hartley and Zisserman [7].

Experiment 1: Calibration from small numbers of correspondences. The first experiment tests the hypothesis that estimating the prior as proposed in this paper allows online camera calibration using only a small number of correspondences. This is important because typical home and office scenes may be relatively devoid of structure and texture, as the examples in figure 1 show.

Our test procedure is as follows. For each set of 10 cameras, the prior is computed using a subset of nine cameras, leaving one camera out in "round-robin" fashion. A small number of points (0–20) is drawn at random without repetition from the correspondence set of the left-out camera, and used to estimate the camera parameters under the computed prior. Then RFE is computed for all (*e.g.* 3000) correspondences for the left-out camera (we take the RMS value). For a given number of points (0–20), we compute median RFE over 50 samples then take the mean over the 10 left-out cameras.

The experiment varies several different factors: parametrization (projective or Euclidean), camera (gluecam or hydra), and prior (datasheet $\boldsymbol{\Sigma}_0$, sample covariance $\bar{\boldsymbol{\Sigma}}$, diagonal sample covariance $\boldsymbol{\Sigma}_d$, and the learned covariance from §3). Finally, the datasheet prior was tested in "broad" and "tight" modes. The tight prior represents a specific prior carefully designed for a particular camera family, as defined in §3, and the broad prior ($\boldsymbol{\Sigma}_0^{\text{broad}} = 10^3\boldsymbol{\Sigma}_0$).

Results are shown in figure 6. Mean RFE in the absence of any correspondence data (*i.e.* using *only* the prior) was 6–8 pixels. For large numbers of correspondences, mean RFE converges to about 0.5 pixels for the Hydra, 0.7 pixels for the gluecam. The results may be summarized as follows. 1. The learned diagonal prior (green curve) generally gives better results than the datasheet prior, with two exceptions: with the carefully constructed Hydra cameras and a tight datasheet prior, performance is equivalent; and with the Euclidean parametrization on the gluecams the learned diagonal prior inhibits bundle adjustment convergence for more than six correspondences. 2. The simple learned priors $\bar{\boldsymbol{\Sigma}}$ (red curve) and $\boldsymbol{\Sigma}_d$ (black curve) do not perform as well as the Bayesian learned prior. 3. For the broad datasheet prior, calibration improves as the number of correspondences approaches four, and then disimproves up to a maximum when the number of correspondences equals seven, the number of degrees of freedom in the system. This is explained because with seven or fewer correspondences, there always exists a parameter vector which yields zero reprojection error. With a broad prior, this estimate is not constrained to be anywhere near the space of likely cameras, and thus the rectification error on the left-out correspondences can be arbitrarily poor. In terms of the epipolar geometries, a broad prior can be as bad as the "no prior" case in figure 1. The learned prior corrects for this effect.

Experiment 2: Real-world scenes. To illustrate the efficacy of the proposed camera calibration algorithm in some typical real world scenes, we used five gluecams to obtain a variety of photographs of some home and office scenes. A representative sample of these photographs is shown in

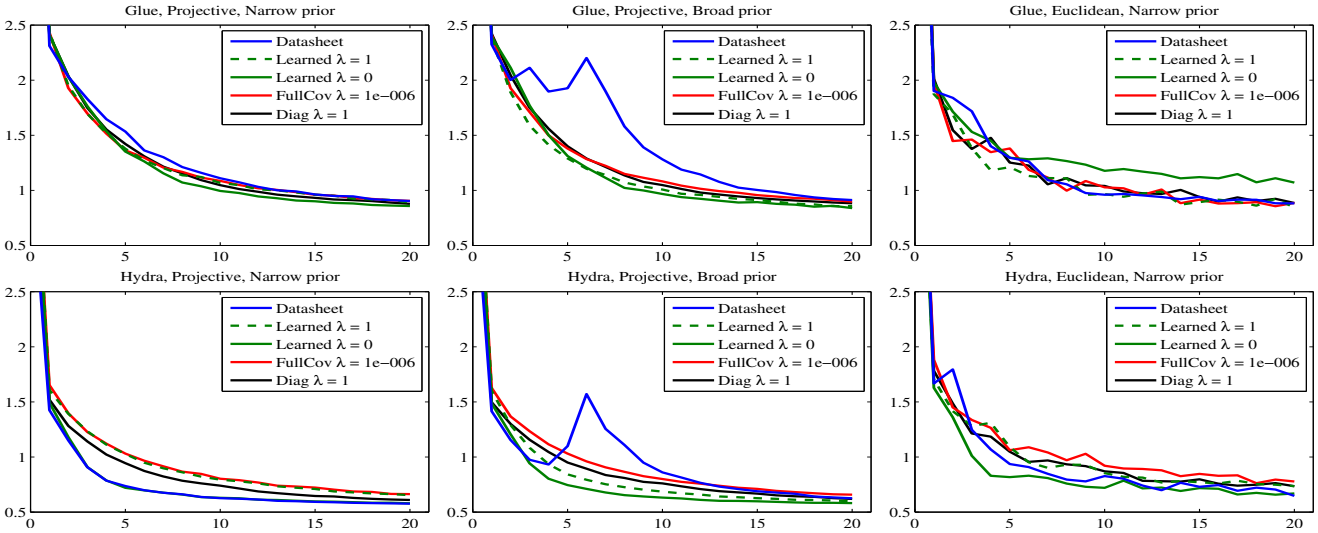

Figure 6. **Calibration accuracy with small numbers of correspondences.** Top row: "gluecam", Bottom row: Hydra. Left: tight datasheet prior, projective parameterization; Middle: broad datasheet prior, projective parameterization; Right: broad datasheet prior, Euclidean parameterization. In the legend, "Learned, $\lambda = 0$" indicates a learned scaled-diagonal prior, "FullCov" corresponds to the regularized sample covariance matrix with fixed regularizer $t = 10^{-6}$, "Learned, $\lambda = 1$" corresponds to a learned regularizer. "Diag" is a scaled diagonal estimate with fixed scale factor λ. Errorbars on all curves are approximately 0.1 pixels.

figure 1. We obtained five photographs per camera, and 25 stereo image pairs in all. Correspondences for each pair of photographs were obtained by detecting Harris corners and performing feature matching using the ground truth camera parameters and approximate knowledge of the range of depths in the scene to constrain the search for correspondences to within a few pixels. Using these correspondences, each camera was calibrated using the prior learned by offline calibration of the remaining nine cameras. Figures 1 and7 demonstrate that accurate and consistent camera calibration is obtained by using the learned prior. Without the prior, conventional calibration fails for all the image pairs in this dataset.

Calibration accuracy was also assessed empirically using the correspondences obtained during offline calibration of the test camera. Mean RFE for the tight datasheet prior was 4.1 pixels, with 44% of calibrations failing. For the best learned prior, mean RFE was 2.8 pixels, and the failure rate reduced to 12%. This is a challenging test because the test correspondences were obtained from photographs of the calibration target located in four positions throughout the entire working volume of the camera—whereas the correspondences used for online calibration were often on a single plane. Nevertheless, online calibration with the learned prior gives useful accuracy across a large working volume a high proportion of the time.

5. Conclusion

Online camera calibration has previously made the hidden assumption of rich scene structure, which is not valid in many application domains, for example home and office environments. This paper has shown that scene correspondences in these environments can be used for calibration, but only if a strong prior is available. Such a prior can be constructed from a small number of examples if care is taken with choice of parameterization and the prior model. Although the subsequent estimation problem is simply an application of standard Bayesian methodology, the contribution of this paper is in identifying the problem, and in suggesting a practical method to obtain the prior. The results show that a correctly learned prior generally does better than a diagonal prior, and that only the learned prior generally outperform calibration with a "blueprint" prior, or with no prior at all.

The optimization for the Bayesian estimate is still somewhat ad-hoc. By restricting to a single-parameter family of covariance matrices we make finding the global optimum easy, but it may be that better priors are to be found using a more general optimization. Our approach may be seen as a variant of approaches to estimating covariance matrices from small datasets in computational biology [15], and we plan to explore this link.

Future work might focus on the automatic identification of camera families from image context. It might be that home and office scenes require different models, which can be identified using object recognition strategies [11]. This paper has concentrated on a single model for all variations, both those in manufacturing and those occurring over time. It would be interesting to see if the parameter space has natural foliations corresponding to these different variations, so that repeated calibration of the same camera could be re-

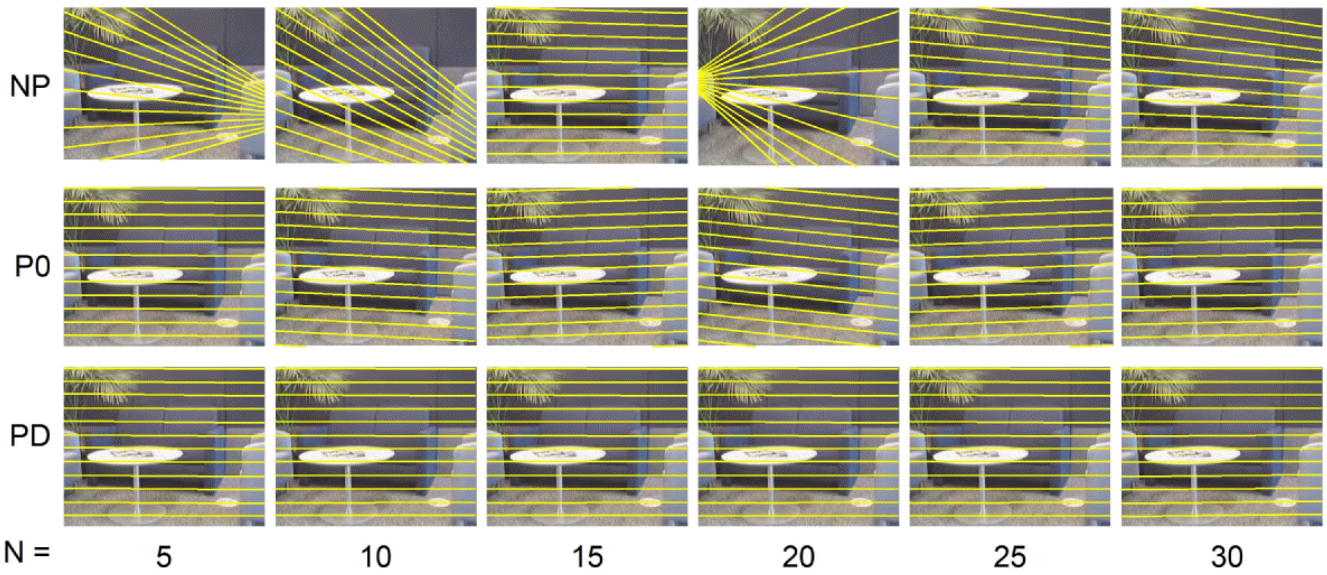

Figure 7. **Stereo camera calibration results in an office scene.** 30 point matches were obtained by hand and the camera was calibrated using random samples of 5, 10, 15, 20, 25, and 30 matches. Recovered epipolar geometry is shown (right eye only) for calibration with no prior (NP), the datasheet prior (P0) and the learned diagonal prior (PD). The ground truth epipolar lines were approximately horizontal.

stricted to correcting for time-varying effects.

References

[1] A. Azarbayejani and A. Pentland. Recursive estimation of motion, structure, and focal length. *PAMI*, 17:562–575, 1995.

[2] C. Bishop. *Pattern recognition and machine learning*. Springer, 2006.

[3] T. Broida, S. Chandrashekhar, and R. Chellappa. Recursive estimation of 3D kinematics and structure from a noisy monocular image sequence. *IEEE Trans. Aerospace Electr. Syst.*, 26:639–656, 1990.

[4] R. A. Brooks, A. N. Flynn, and T. Marill. Self calibration of motion and stereo vision for mobile robots. In *Proc. Fourth Intl Symp. Robotics Research*, 1987.

[5] A. J. Davison. Real-time simultaneous localisation and mapping with a single camera. In *ICCV*, pages 1403–1410, 2003.

[6] M. Grossberg and S. Nayar. A general imaging model and a method for finding its parameters. In *ICCV*, volume 2, pages 108–115, 2001.

[7] R. Hartley and A. Zisserman. *Multiple view geometry in computer vision*. Cambridge University Press, 2000.

[8] HYtek Automation. 3D iVCam—Stereo 3D Webcam Suite, 2007. http://www.hytekautomation.com/Products/Stereo-Webcam.html, Viewed April 2007.

[9] P. McLauchlan and D. Murray. Active camera calibration for a head-eye platform using the variable state-dimension filter. *PAMI*, 18(1):15–22, 1996.

[10] nVela. Hydra—Stereo Webcam, 2007. http://www.nvela.com, Viewed August 2007.

[11] A. Oliva and A. Torralba. Building the gist of a scene: The role of global image features in recognition. *Visual Perception, Progress in Brain Research*, 155, 2006.

[12] Point Grey Research. Bumblebee, 2007. http://www. ptgrey.com/ products/ bumblebee2/ index.asp, website viewed April 2007.

[13] M. Pollefeys, F. Verbiest, and L. J. V. Gool. Surviving dominant planes in uncalibrated structure and motion recovery. In *ECCV (2)*, pages 837–851, 2002.

[14] G. Qian and R. Chellappa. Bayesian self-calibration of a moving camera. *Computer Vision and Image Understanding*, 95:287–316, 2004.

[15] J. Schäfer and K. Strimmer. A shrinkage approach to large-scale covariance matrix estimation and implications for functional genomics. *Statistical Applications in Genetics and Molecular Biology*, 4(1), 2005.

[16] R. J. Schalkoff. Automatic recalibration of moving cameras in stereo vision systems. *Image and vision computing*, 3(3):118–121, 1985.

[17] C. Slama. *Manual of Photogrammetry*. American Society of Photogrammetry, Falls Church, VA, USA, 4th edition, 1980.

[18] P. Sturm and S. Ramalingam. A generic concept for camera calibration. In *ECCV*, volume 2, pages 1–13, 2004.

[19] T. Svoboda, D. Martinec, and T. Pajdla. A convenient multi-camera self-calibration for virtual environments. *PRESENCE: Teleop. and Virtual Env.*, 14(4):407–422, 2005.

[20] P. H. S. Torr, A. W. Fitzgibbon, and A. Zisserman. Maintaining multiple motion model hypotheses over many views. In *ICCV*, pages 485–491, 1998.

[21] B. Triggs, P. McLauchlan, R. Hartley, and A. Fitzgibbon. Bundle adjustment: A modern synthesis. In *Vision algorithms: Theory and practice, Springer LNCS 1883*, 2000.