# On Calibration, Structure-From-Motion and Multi-View Geometry for General Camera Models

Peter Sturm, Srikumar Ramalingam, Suresh K. Lodha

# ON CALIBRATION, STRUCTURE-FROM-MOTION AND MULTI-VIEW GEOMETRY FOR PANORAMIC CAMERA MODELS

Peter Sturm[a], Srikumar Ramalingam[b], Suresh K. Lodha[b]

[a] INRIA Rhône-Alpes, 655 Avenue de l'Europe, 38330 Montbonnot, France – Peter.Sturm@inrialpes.fr
[b] Dept. of Computer Science, University of California, Santa Cruz, USA – {srikumar,lodha}@cse.ucsc.edu

**Commission V, WG V/1 and 5**

**KEY WORDS:** Panoramic camera, Non-central camera, Multi-view geometry, Calibration, 3D Reconstruction, Motion estimation

**ABSTRACT**

We consider calibration and structure-from-motion tasks for a previously introduced, highly general imaging model, where cameras are modeled as possibly unconstrained sets of projection rays. This allows to describe most existing camera types (at least for those operating in the visible domain), including pinhole cameras, sensors with radial or more general distortions, and especially panoramic cameras (central or non-central). Generic algorithms for calibration and structure-from-motion tasks (absolute and relative orientation, 3D point triangulation) are outlined. The foundation for a multi-view geometry of non-central cameras is given, leading to the formulation of multi-view matching tensors, analogous to the essential matrix, trifocal and quadrifocal tensors of perspective cameras. Besides this, we also introduce a natural hierarchy of camera models: the most general model has unconstrained projection rays whereas the most constrained model dealt with here is the central one, where all rays pass through a single point.

## 1 INTRODUCTION

Many different types of cameras including pinhole, stereo, catadioptric, omnidirectional and non-central cameras have been used in computer vision and photogrammetry. Most existing camera models are parametric (i.e. defined by a few intrinsic parameters) and address imaging systems with a single effective viewpoint (all rays pass through one point). In addition, existing calibration or structure-from-motion procedures are often taylor-made for specific camera models, see examples e.g. in (Barreto & Araujo, 2003; Gruen & Huang, 2001; Hartley & Zisserman, 2000; Geyer & Daniilidis, 2002).

The aim of this work is to relax these constraints: we want to propose and develop calibration and structure-from-motion methods that should work for any type of camera model, and especially also for cameras without a single effective viewpoint. To do so, we first renounce on parametric models, and adopt the following very general model: a camera acquires images consisting of pixels; each pixel captures light that travels along a ray in 3D. The camera is fully described by (Grossberg & Nayar, 2001):

- the coordinates of these rays (in a local coordinate frame).
- the mapping between rays and pixels; this is basically a simple indexing.

This is of course an idealistic model; other aspects, e.g. photometry and point-spread function are described in (Grossberg & Nayar, 2001). This general imaging model allows to describe virtually any camera that captures light rays travelling along straight lines. Examples are (cf. figure 1):

- a camera with any type of optical distortion, e.g. radial or tangential.
- a camera looking at a reflective surface, e.g. as often used in surveillance, a camera looking at a spherical or otherwise curved mirror (Hicks & Bajcsy, 2000). Such systems, as opposed to central catadioptric devices using parabolic or hyperbolic mirrors (Baker & Nayar, 1999; Geyer & Daniilidis, 2000), do not usually have a single effective viewpoint.
- multi-camera stereo systems: put together the pixels of all image planes; they "catch" light rays that do not travel along lines that all pass through a single point. Nevertheless, in the above general camera model, a stereo system (with rigidly linked cameras) is considered as a **single** camera.

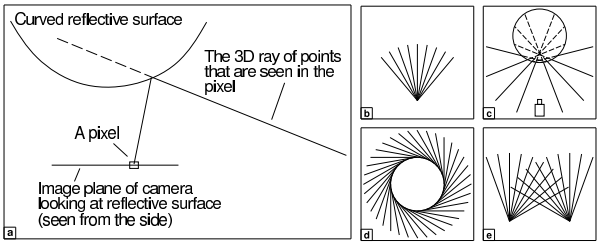


Figure 1: Examples of imaging systems. (a) Catadioptric system. Note that camera rays do not pass through their associated pixels. (b) Central camera (e.g. perspective, with or without radial distortion). (c) Camera looking at reflective sphere. This is a non-central device (camera rays are not intersecting in a single point). (d) Omnivergent imaging system (Peleg 2001; Shum 1999). (e) Stereo system (non-central) consisting of two central cameras.

- other acquisition systems, many of them being non-central, see e.g. (Bakstein, 2001; Bakstein & Pajdla, 2001; Neuman et al., 2003; Pajdla, 2002b; Peleg et al., 2001; Shum et al., 1999; Swaminathan et al., 2003; Yu & McMillan, 2004), insect eyes, etc.

In this article, we first review some recent work on calibration and structure-from-motion for this general camera model. Concretely, we outline basics for calibration, pose and motion estimation, as well as 3D point triangulation. We then describe the foundations for a mult-view geometry of the general, non-central camera model, leading to the formulation of multi-view matching tensors, analogous to the fundamental matrices, trifocal and quadrifocal tensors of perspective cameras. Besides this, we also introduce a natural hierarchy of camera models: the most general model has unconstrained projection rays whereas the most constrained model dealt with here is the central model, where all rays pass through a single point. An intermediate model is what we term *axial cameras*: cameras for which there exists a 3D line that cuts all projection rays. This encompasses for example x-slit projections, linear pushbroom cameras and some non-central catadioptric systems. Hints will be given how to adopt the multi-view geometry proposed for the general imaging model, to such axial cameras.

The paper is organized as follows. A hierarchy of camera models is proposed in section 2. Sections 3 to 5 deal with calibration, pose estimation, motion estimation, as well as 3D point triangulation. The multi-view geometry for the general camera model is given in section 6. A few experimental results are shown in section 7.

## 2 CAMERA MODELS

A **non-central camera** may have completely unconstrained projection rays, whereas for a **central camera**, there exists a point – the **optical center** – that lies on all projection rays. An intermediate case is what we call **axial cameras**, where there exists a line that cuts all projection rays – the **camera axis** (not to be confounded with optical axis). Examples of cameras falling into this class are:

- x-slit cameras (Pajdla, 2002a; Zomet et al., 2003) (also called two-slit or crossed-slits cameras), and their special case of linear pushbroom cameras (Hartley & Gupta, 1994). Note that these form a sub-class of axial cameras, see below.
- stereo systems consisting of 2 central cameras or 3 or more central cameras with collinear optical centers.
- non-central catadioptric cameras of the following construction: the mirror is any surface of revolution and the optical center of the central camera (can be any central camera, i.e. not necessarily a pinhole) looking at the mirror lies on its axis of revolution. It is easy to verify that in this case, all projection rays cut the mirror's axis of revolution, i.e. the camera is an axial camera, with the mirror's axis of revolution as camera axis. Note that catadioptric cameras with a spherical mirror and a central camera looking at it, are always non-central, and are actually always axial cameras.

These three classes of camera models may also be defined as: existence of a linear space of $d$ dimensions that has an intersection with all projection rays. In this sense, $d = 0$ defines central cameras, $d = 1$ axial cameras and $d = 2$ general non-central cameras.

Intermediate classes do exist. X-slit cameras are a special case of axial cameras: there actually exist 2 lines in space that both cut all projection rays. Similarly, central 1D cameras (cameras with a single row of pixels) can be defined by a point and a line in 3D. Camera models, some of which do not have much practical importance, are summarized in table 1. A similar way of defining camera types was suggested in (Pajdla, 2002a).

It is worthwhile to consider different classes due to the following observation: the usual calibration and motion estimation algorithms proceed by first estimating a matrix or tensor by solving linear equation systems (e.g. the calibration tensors in (Sturm & Ramalingam, 2004) or the essential matrix (Pless, 2003)). Then, the parameters that are searched for (usually, motion parameters), are extracted from these. However, when estimating for example the $6 \times 6$ essential matrix of *non-central* cameras based on image correspondences obtained from *central* or *axial* cameras, then the associated linear equation system does not give a unique solution. Consequently, the algorithms for extracting the actual motion parameters, can not be applied without modification.

| Points/lines cutting rays | Description |
|---|---|
| None | Non-central camera |
| 1 point | Central camera |
| 2 points | Camera with a single projection ray |
| 1 line | Axial camera |
| 1 point, 1 line | Central 1D camera |
| 2 skew lines | X-slit camera |
| 2 coplanar lines | Union of a non-central 1D camera and a central camera |
| 3 coplanar lines without a common point | Non-central 1D camera |

Table 1: Camera models, defined by 3D points and lines that have an intersection with all projection rays of a camera.

## 3 CALIBRATION

### 3.1 Basic Approach

We briefly review a generic calibration approach developed in (Sturm & Ramalingam, 2004), an extension of (Champleboux et al., 1992; Gremban et al, 1988; Grossberg & Nayar, 2001), to calibrate different camera systems. As mentioned, calibration consists in determining, for every pixel, the 3D projection ray associated with it. In (Grossberg & Nayar, 2001), this is done as follows: two images of a calibration object with known structure are taken. We suppose that for every pixel, we can determine the point on the calibration object, that is seen by that pixel[1]. For each pixel in the image, we thus obtain two 3D points. Their coordinates are usually only known in a coordinate frame attached to the calibration object; however, if one knows the motion between the two object positions, one can align the coordinate frames. Then, every pixel's projection ray can be computed by simply joining the two observed 3D points.

In (Sturm & Ramalingam, 2004), we propose a more general approach, that does not require knowledge of the calibration object's displacement. In that case, three images need to be taken at least. The fact that all 3D points observed by a pixel in different views, are on a line in 3D, gives a constraint that allows to recover both the motion and the camera's calibration. The constraint is formulated via a set of trifocal tensors, that can be estimated linearly, and from which motion, and then calibration, can be extracted. In (Sturm & Ramalingam, 2004), this approach is first formulated for the use of 3D calibration objects, and for the general imaging model, i.e. for non-central cameras. We also propose variants of the approach, that may be important in practice: first, due to the usefulness of planar calibration patterns, we specialized the approach appropriately. Second, we propose a variant that works specifically for central cameras (pinhole, central catadioptric, or any other central camera). More details are given in (Sturm & Ramalingam, 2003).

This basic approach only handles the minimum number of images (two respectively three, for central respectively non-central cameras). Also, it only allows to calibrate the pixels that are matched to the calibration object in all images. Especially for panoramic cameras, complete calibration with this approach is thus very hard (unless an "omnidirectional" calibration object is available). Recently, we have thus developed an approach that deals with these drawbacks; it handles any number of images and also allows to calibrate image regions that are not covered by the calibration object in all images. This approach is described in the next paragraph.

### 3.2 General Approach

We propose two ideas to overcome the above mentioned limitations of our basic calibration approach. First, we have recently developed a method along the lines of (Sturm & Ramalingam, 2004) that can use more than the minimum number of images. This method can not be described in full detail here; it will be given in a future publication. This method nevertheless has the drawback of only allowing to calibrate image regions that are covered by the calibration object in all images used.

Our second idea is relatively straightforward. We first perform

[1] This can be achieved for example by using a flat screen as calibration "grid" and taking images of several black & white patterns that together uniquely encode the position of pixels on the screen.

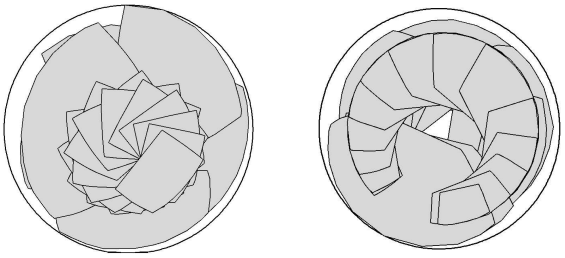

Figure 2: Examples of image regions corresponding to different images of calibration objects. Left: 23 images of calibration objects with a fisheye camera. Right: 24 images with a spherical catadioptric camera.

an initial calibration using our basic approach. This only allows to calibrate an image region that is covered by the calibration object in all images used. We then extend the calibration to the rest of the image, as follows. For each image in which the calibration object covers a sufficiently large already calibrated region, we can compute the object's pose relative to the camera (see section 4.1). Then, for each as yet uncalibrated pixel, we check if it is matched to the calibration object in sufficiently many images (one for central cameras, two for non-central ones); if so, we can compute the coordinates of its projection ray. For a non-central camera, we simply fit a straight line to the matching 3D points on the calibration object for different positions/images. As for the central model, we compute a straight line that is constrained to pass through the optical center.

These two procedures – computation of pose and projection rays – are repeated in alternation, until all available images have been used. Figure 2 gives examples of image regions covered by calibration objects in different images, for panoramic cameras that have been calibrated using our approach.

We also have developed a bundle adjustment that can be used between iterations, or only at the end of the above process, to refine calibration and pose. Our bundle adjustment minimizes *ray–point* distance, i.e. the distance in 3D, between projection rays and matching points on calibration objects. This is not the optimal measure, but reprojection-based bundle adjustment is not trivial to formulate for the generic imaging model (some ideas on this are given in (Ramalingam et al., 2004)). The minimization is done for the optical center position (only for central cameras), the pose of calibration objects, and of course the coordinates of projection rays. The ray–point distance is computed as

$$E = \sum_{i=1}^{r} \sum_{j=1}^{n} \|\mathbf{C}_i + \lambda_{ij}\mathbf{D}_i - \mathsf{R}_j\mathbf{P}_{ij} - \mathbf{t}_j\|^2$$

with:

- $n$ is the number of calibration objects and $r$ the number of rays.
- $\mathbf{C}_i$ is a point on the $i$th ray (in the non-central case) or the optical center (in a central model).
- $\mathbf{D}_i$ is the direction of the $i$th ray.
- $\lambda_{ij}$ parameterizes the point on the $i$th ray that should correspond to its intersection with the $j$th calibration object.
- $\mathbf{P}_{ij}$ is the point on the $j$th calibration object that is matched to the pixel associated with the $i$th ray.
- $\mathsf{R}_j$ and $\mathbf{t}_j$ represent the pose of the $j$th calibration object.

## 4 ORIENTATION

### 4.1 Pose Estimation

Pose estimation is the problem of computing the relative position and orientation between an object of *known* structure, and a calibrated camera. A literature review on algorithms for pinhole cameras is given in (Haralick et al., 1994). Here, we briefly show how the minimal case can be solved for general cameras. For pinhole cameras, pose can be estimated, up to a finite number of solutions, from 3 point correspondences (3D-2D) already. The same holds for general cameras. Consider 3 image points and the associated projection rays, computed using the calibration information. We parameterize generic points on the rays as follows: $\mathbf{A}_i + \lambda_i\mathbf{B}_i$.

We know the structure of the observed object, meaning that we know the mutual distances $d_{ij}$ between the 3D points. We can thus write equations on the unknowns $\lambda_i$, that parameterize the object's pose:

$$\|\mathbf{A}_i + \lambda_i\mathbf{B}_i - \mathbf{A}_j - \lambda_j\mathbf{B}_j\|^2 = d_{ij}^2 \quad \text{for } (i,j) = (1,2),(1,3),(2,3)$$

This gives a total of 3 equations that are quadratic in 3 unknowns. Many methods exist for solving this problem, e.g. symbolic computation packages such as MAPLE allow to compute a resultant polynomial of degree 8 in a single unknown, that can be numerically solved using any root finding method.

Like for pinhole cameras, there are up to 8 theoretical solutions. For pinhole cameras, at least 4 of them can be eliminated because they would correspond to points lying behind the camera (Haralick et al., 1994). As for general cameras, determining the maximum number of feasible solutions requires further investigation. In any case, a unique solution can be obtained using one or two additional points (Haralick et al., 1994). More details on pose estimation for non-central cameras are given in (Chen & Chang, 2004; Nistér, 2004).

### 4.2 Motion Estimation

We outline how ego-motion, or, more generally, relative position and orientation of two calibrated general cameras, can be estimated. This is done via a generalization of the classical motion estimation problem for pinhole cameras and its associated centerpiece, the essential matrix (Longuet-Higgins, 1981). We briefly summarize how the classical problem is usually solved (Hartley & Zisserman, 2000). Let $\mathsf{R}$ be the rotation matrix and $\mathbf{t}$ the translation vector describing the motion. The essential matrix is defined as $\mathsf{E} = -[\mathbf{t}]_\times\mathsf{R}$. It can be estimated using point correspondences $(\mathbf{x}_1, \mathbf{x}_2)$ across two views, using the epipolar constraint $\mathbf{x}_2^\mathsf{T}\mathsf{E}\mathbf{x}_1 = 0$. This can be done linearly using 8 correspondences or more. In the minimal case of 5 correspondences, an efficient non-linear minimal algorithm, which gives exactly the theoretical maximum of 10 feasible solutions, was only recently introduced (Nistér, 2003). Once the essential matrix is estimated, the motion parameters $\mathsf{R}$ and $\mathbf{t}$ can be extracted relatively straightforwardly (Nistér, 2003).

In the case of our general imaging model, motion estimation is performed similarly, using pixel correspondences $(\mathbf{x}_1, \mathbf{x}_2)$. Using the calibration information, the associated projection rays can be computed. Let them be represented by their Plücker coordinates (see section 6), i.e. 6-vectors $\mathbf{L}_1$ and $\mathbf{L}_2$. The epipolar constraint extends naturally to rays, and manifests itself by a $6 \times 6$ essential matrix (Pless, 2003):

$$\mathsf{E} = \begin{pmatrix} -[\mathbf{t}]_\times\mathsf{R} & \mathsf{R} \\ \mathsf{R} & 0 \end{pmatrix}$$

The epipolar constraint then writes: $\mathbf{L}_2^\mathsf{T}\mathsf{E}\mathbf{L}_1 = 0$ (Pless, 2003). Once $\mathsf{E}$ is estimated, motion can again be extracted straightforwardly (e.g., $\mathsf{R}$ can simply be read off $\mathsf{E}$). Linear estimation of $\mathsf{E}$ requires 17 correspondences.

There is an important difference between motion estimation for central and non-central cameras: with central cameras, the translation component can only be recovered up to scale. Non-central

cameras however, allow to determine even the translation's scale. This is because a single calibrated non-central camera already carries scale information (via the distance between mutually skew projection rays). One consequence is that the theoretical minimum number of required correspondences is 6 instead of 5. It might be possible, though very involved, to derive a minimal 6-point method along the lines of (Nistér, 2003).

More details on motion estimation for non-central cameras and intermediate camera models, will be given in a forthcoming publication.

## 5 3D RECONSTRUCTION

We now describe an algorithm for 3D reconstruction from two or more calibrated images with known relative position. Let $\mathbf{C} = (X, Y, Z)^\mathsf{T}$ be a 3D point that is to be reconstructed, based on its projections in $n$ images. Using calibration information, we can compute the $n$ associated projection rays. Here, we represent the $i$th ray using a starting point $\mathbf{A}_i$ and the direction, represented by a unit vector $\mathbf{B}_i$. We apply the mid-point method (Hartley & Sturm, 1997; Pless, 2003), i.e. determine $\mathbf{C}$ that is closest in average to the $n$ rays. Let us represent generic points on rays using position parameters $\lambda_i$, as in the previous section. Then, $\mathbf{C}$ is determined by minimizing the following expression over $\mathbf{C}^\mathsf{T} = (X, Y, Z)$ and the $\lambda_i$: $\sum_{i=1}^{n} \|\mathbf{A}_i + \lambda_i \mathbf{B}_i - \mathbf{C}\|^2$.

This is a linear least squares problem, which can be solved e.g. via the Pseudo-Inverse, leading to the following explicit equation (derivations omitted):

$$\begin{pmatrix} \mathbf{C} \\ \lambda_1 \\ \vdots \\ \lambda_n \end{pmatrix} = \mathsf{M}^{-1} \begin{pmatrix} \mathtt{I}_3 & \cdots & \mathtt{I}_3 \\ -\mathbf{B}_1^\mathsf{T} & & \\ & \ddots & \\ & & -\mathbf{B}_n^\mathsf{T} \end{pmatrix} \begin{pmatrix} \mathbf{A}_1 \\ \vdots \\ \mathbf{A}_n \end{pmatrix}$$

with

$$\mathsf{M} = \begin{pmatrix} n\mathtt{I}_3 & -\mathbf{B}_1 & \cdots & -\mathbf{B}_n \\ -\mathbf{B}_1^\mathsf{T} & 1 & & \\ \vdots & & \ddots & \\ -\mathbf{B}_n^\mathsf{T} & & & 1 \end{pmatrix}$$

where $\mathtt{I}_3$ is the identity matrix of size $3 \times 3$. Due to its sparse structure, the inversion of $\mathsf{M}$ can actually be performed in closed-form. Overall, the triangulation of a 3D point using $n$ rays, can by carried out very efficiently, using only matrix multiplications and the inversion of a symmetric $3 \times 3$ matrix.

## 6 MULTI-VIEW GEOMETRY

We establish the foundations of a multi-view geometry for general (non-central) cameras. Its cornerstones are, as with perspective cameras, matching tensors. We show how to establish them, analogously to the perspective case.

Here, we only talk about the calibrated case; the uncalibrated case is nicely treated for perspective cameras, since calibrated and uncalibrated cameras are linked by projective transformations. For non-central cameras however, there is no such link: in the most general case, every pair (pixel, camera ray) may be completely independent of other pairs.

### 6.1 Reminder on Multi-View Geometry for Perspective Cameras

We briefly review how to derive multi-view matching relations for perspective cameras (Faugeras & Mourrain, 1995). Let $\mathsf{P}_i$ be projection matrices and $\mathbf{q}_i$ image points. A set of image points are matching, if there exists a 3D point $\mathbf{Q}$ and scale factors $\lambda_i$ such that:

$$\lambda_i \mathbf{q}_i = \mathsf{P}_i \mathbf{Q}$$

This may be formulated as the following matrix equation:

$$\underbrace{\begin{pmatrix} \mathsf{P}_1 & \mathbf{q}_1 & \mathbf{0} & \cdots & \mathbf{0} \\ \mathsf{P}_2 & \mathbf{0} & \mathbf{q}_2 & \cdots & \mathbf{0} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \mathsf{P}_n & \mathbf{0} & \mathbf{0} & \cdots & \mathbf{q}_n \end{pmatrix}}_{\mathsf{M}} \begin{pmatrix} \mathbf{Q} \\ -\lambda_1 \\ -\lambda_2 \\ \vdots \\ -\lambda_n \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{pmatrix}$$

The matrix $\mathsf{M}$, of size $3n \times (4+n)$ has thus a null-vector, meaning that its rank is less than $4+n$. Hence, the determinants of all its submatrices of size $(4+n)\times(4+n)$ must vanish. These determinants are multi-linear expressions in terms of the coordinates of image points $\mathbf{q}_i$.

They have to be considered for every possible submatrix. Only submatrices with 2 or more rows per view, give rise to constraints linking all projection matrices. Hence, constraints can be obtained for up to $n$ views with $2n \leq 4+n$, meaning that only for up to 4 views, matching constraints linking all views can be obtained.

The constraints for $n$ views take the form:

$$\sum_{i_1=1}^{3} \sum_{i_2=1}^{3} \cdots \sum_{i_n=1}^{3} q_{1,i_1} q_{2,i_2} \cdots q_{n,i_n} T_{i_1,i_2,\cdots,i_n} = 0 \qquad (1)$$

where the multi-view matching tensor $\mathsf{T}$ of dimension $3 \times \cdots \times 3$ depends on and partially encodes the cameras' projection matrices $\mathsf{P}_i$. Note that as soon as cameras are calibrated, this theory applies to any central camera: for a camera with radial distortion for example, the above formulation holds for distortion-corrected image points.

### 6.2 Multi-View Geometry for Non-Central Cameras

Here, instead of projection matrices (depending on calibration and pose), we deal with pose matrices:

$$\mathsf{P}_i = \begin{pmatrix} \mathsf{R}_i & \mathbf{t}_i \\ \mathbf{0}^\mathsf{T} & 1 \end{pmatrix}$$

These express the similarity transformations that map a point from some global reference frame, into the cameras' local coordinate frames (since no optical center and no camera axis exist, no assumptions about the local coordinate frames are made). As for image points, they are now replaced by camera rays. Let the $i$th ray be represented by two 3D points $\mathbf{A}_i$ and $\mathbf{B}_i$. Eventually, we will to obtain expressions in terms of the rays' Plücker coordinates. Plücker coordinates can be defined in various ways; the definition we use is as follows. The line can be represented by the skew-symmetric $4 \times 4$ so-called Plücker matrix

$$\mathsf{L} = \mathbf{A}\mathbf{B}^\mathsf{T} - \mathbf{B}\mathbf{A}^\mathsf{T}$$

Note that the Plücker matrix is independent (up to scale) of which pair of points on the line are chosen to represent it. An alternative representation for the line is its Plücker coordinate vector of length 6:

$$\mathbf{L} = \begin{pmatrix} A_4B_1 - A_1B_4 \\ A_4B_2 - A_2B_4 \\ A_4B_3 - A_3B_4 \\ A_3B_2 - A_2B_3 \\ A_1B_3 - A_3B_1 \\ A_2B_1 - A_1B_2 \end{pmatrix} \qquad (2)$$

Our goal is to obtain matching tensors $\mathsf{T}$ and matching constraints of the form (1), with the difference that tensors will have size $6 \times \cdots \times 6$ and act on Plücker line coordinates:

$$\sum_{i_1=1}^{6} \sum_{i_2=1}^{6} \cdots \sum_{i_n=1}^{6} L_{1,i_1} L_{2,i_2} \cdots L_{n,i_n} T_{i_1,i_2,\cdots,i_n} = 0 \qquad (3)$$

In the following, we explain how to derive such matching constraints. Consider a set of $n$ camera rays and let them be defined by two points $\mathbf{A}_i$ and $\mathbf{B}_i$ each; the choice of points to represent a ray is not important, since later we will fall back onto the ray's Plücker coordinates.

Now, a set of $n$ camera rays are matching, if there exist a 3D point $\mathbf{Q}$ and scale factors $\lambda_i$ and $\mu_i$ associated with each ray such that:

$$\lambda_i \mathbf{A}_i + \mu_i \mathbf{B}_i = \mathsf{P}_i \mathbf{Q}$$

i.e. if the point $\mathsf{P}_i\mathbf{Q}$ lies on the line spanned by $\mathbf{A}_i$ and $\mathbf{B}_i$. As for perspective cameras, we group these equations in matrix form:

$$\mathsf{M} \begin{pmatrix} \mathbf{Q} \\ -\lambda_1 \\ -\mu_1 \\ -\lambda_2 \\ -\mu_2 \\ \vdots \\ -\lambda_n \\ -\mu_n \end{pmatrix} = \begin{pmatrix} \mathbf{0} \\ \mathbf{0} \\ \vdots \\ \mathbf{0} \end{pmatrix}$$

with:

$$\mathsf{M} = \begin{pmatrix} \mathsf{P}_1 & \mathbf{A}_1 & \mathbf{B}_1 & \mathbf{0} & \mathbf{0} & \cdots & \mathbf{0} & \mathbf{0} \\ \mathsf{P}_2 & \mathbf{0} & \mathbf{0} & \mathbf{A}_2 & \mathbf{B}_2 & \cdots & \mathbf{0} & \mathbf{0} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ \mathsf{P}_n & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \cdots & \mathbf{A}_n & \mathbf{B}_n \end{pmatrix}$$

As above, this equation shows that $\mathsf{M}$ must be rank-deficient. However, the situation is different here since the $\mathsf{P}_i$ are of size $4 \times 4$ now, and $\mathsf{M}$ of size $4n \times (4+2n)$. We thus have to consider submatrices of $\mathsf{M}$ of size $(4+2n) \times (4+2n)$. Furthermore, in the following we show that only submatrices with 3 rows or more per view, give rise to constraints on all pose matrices. Hence, $3n \leq 4 + 2n$, and again, $n \leq 4$, i.e. multi-view constraints are only obtained for up to 4 views.

Let us first see what happens for a submatrix of $\mathsf{M}$ where some view contributes only a single row. The two columns corresponding to its base points $\mathbf{A}$ and $\mathbf{B}$, are multiples of one another since they consist of zeroes only, besides a single non-zero coefficient, in the single row associated with the considered view. Hence, the determinant of the considered submatrix of $\mathsf{M}$ is always zero, and no constraint is available.

In the following, we exclude this case, i.e. we only consider submatrices of $\mathsf{M}$ where each view contributes at least 2 rows. Let $\mathsf{N}$ be such a matrix. Without loss of generality, we start to develop its determinant with the columns containing $\mathbf{A}_1$ and $\mathbf{B}_1$. The determinant is then given as a sum of terms of the form:

$$(A_{1,j}B_{1,k} - A_{1,k}B_{1,j}) \det \bar{\mathsf{N}}_{jk}$$

where $j, k \in \{1..4\}$, $j \neq k$, and $\bar{\mathsf{N}}_{jk}$ is obtained from $\mathsf{N}$ by dropping the columns containing $\mathbf{A}_1$ and $\mathbf{B}_1$ as well as the rows containing $A_{1,j}$ etc.

We observe several things:

- The term $(A_{1,j}B_{1,k} - A_{1,k}B_{1,j})$ is nothing else than one of the Plücker coordinates of the ray of camera 1 (cf. (2)). By continuing with the development of the determinant of $\bar{\mathsf{N}}_{jk}$, it becomes clear that the total determinant of $\mathsf{N}$ can be written in the form:

  $$\sum_{i_1=1}^{6} \sum_{i_2=1}^{6} \cdots \sum_{i_n=1}^{6} L_{1,i_1} L_{2,i_2} \cdots L_{n,i_n} T_{i_1,i_2,\cdots,i_n} = 0$$

  i.e. the coefficients of the $\mathbf{A}_i$ and $\mathbf{B}_i$ are "folded together" into the Plücker coordinates of camera rays and $\mathsf{T}$ is a matching tensor between the $n$ cameras. Its coefficients depend exactly on the cameras' pose matrices.

| | central | | non-central | |
|---|---|---|---|---|
| # views | $\mathsf{M}$ | useful | $\mathsf{M}$ | useful |
| 2 | $6 \times 6$ | 3-3 | $8 \times 8$ | 4-4 |
| 3 | $9 \times 7$ | 3-2-2 | $12 \times 10$ | 4-3-3 |
| 4 | $12 \times 8$ | 2-2-2-2 | $16 \times 12$ | 3-3-3-3 |

Table 2: Cases of multi-view matching constraints for central and non-central cameras. The columns entitled "useful" contain entries of the form $x - y - z$ etc. that correspond to sub-matrices of $\mathsf{M}$ that give rise to matching constraints linking *all* views: $x - y - z$ etc. refers to submatrices of $\mathsf{M}$ containing $x$ rows from one camera, $y$ from another etc.

- If camera 1 contributes only two rows to $\mathsf{N}$, then the determinant of $\mathsf{N}$ becomes of the form:

  $$L_{1,x} \left( \sum_{i_2=1}^{6} \cdots \sum_{i_n=1}^{6} L_{2,i_2} \cdots L_{n,i_n} T_{i_2,\cdots,i_n} \right) = 0$$

  i.e. it only contains a single coordinate of the ray of camera 1, and the tensor $\mathsf{T}$ does not depend at all on the pose of that camera. Hence, to obtain constraints between all cameras, every camera has to contribute at least three rows to the considered submatrix.

We are now ready to establish the different cases that lead to useful multi-view constraints. As mentioned above, for more than 4 cameras, no constraints linking all of them are available: submatrices of size at least $3n \times 3n$ would be needed, but $\mathsf{M}$ only has $4 + 2n$ columns. So, only for $n \leq 4$, such submatrices exist.

Table 2 gives all useful cases, both for central and non-central cameras. These lead to two-view, three-view and four-view matching constraints, encoded by essential matrices, trifocal and quadrifocal tensors. Deriving their forms is now mainly a mechanical task.

### 6.3 Multi-View Geometry for Intermediate Camera Models

This multi-view geometry can be specialized to some of the intermediate camera models described in section 2. We have derived this for the axial and x-slit camera models. This will be reported elsewhere in detail.

## 7 EXPERIMENTAL RESULTS

We have calibrated a wide variety of cameras (both central and non-central) as shown in Table 3. Results are first discussed for several "slightly non-central" cameras and for a multi-camera system. We then report results for structure-from-motion algorithms, applied to setups combining cameras of different types (pinhole and panoramic).

**Slightly non-central cameras: central vs. non-central models.** For three cameras (a fisheye, a hyperbolic and a spherical catadioptric system, see sample images in Figure 3), we applied our calibration approach with both, a central and a non-central model assumption. Table 3 shows that the bundle adjustment's residual errors for central and non-central calibration, are very close to one another for the fisheye and hyperbolic catadioptric cameras. This suggests that for the cameras used in the experiments, the central model is appropriate. As for the spherical catadioptric camera, the non-central model has a significantly lower residual, which may suggest that a non-central model is better here.

To further investigate this issue we performed another evaluation. A calibration grid was put on a turntable, and images were acquired for different turntable positions. We are thus able to quantitatively evaluate the calibration, by measuring how close the recovered grid pose corresponds to a turntable sequence. Individual grid points move on a circle in 3D; we thus compute a least squares circle fit to the 3D positions given by the estimated grid

| Camera | Images | Rays | Points | RMS |
|---|---|---|---|---|
| Pinhole (C) | 3 | 217 | 651 | 0.04 |
| Fisheye (C) | 23 | 508 | 2314 | 0.12 |
| (NC) | 23 | 342 | 1712 | 0.10 |
| Sphere (C) | 24 | 380 | 1441 | 2.94 |
| (NC) | 24 | 447 | 1726 | 0.37 |
| Hyperbolic (C) | 24 | 293 | 1020 | 0.40 |
| (NC) | 24 | 190 | 821 | 0.34 |
| Multi-Cam (NC) | 3 | 1156 | 3468 | 0.69 |
| Eye+Pinhole (C) | 3 | 29 | 57 | 0.98 |

Table 3: Bundle adjustment statistics for different cameras. (C) and (NC) refer to central and non-central calibration respectively, and RMS is the root-mean-square residual error of the bundle adjustment (ray-point distances). It is given in percent, relative to the overall size of the scene (largest pairwise distance between points on calibration grids).

| Camera | Grids | Central | Non-Central |
|---|---|---|---|
| Fisheye | 14 | 0.64 | 0.49 |
| Spherical | 19 | 2.40 | 1.60 |
| Hyperbolic | 12 | 0.81 | 1.17 |

Table 4: RMS error for circle fits to grid points, for turntable sequences (see text).

pose. At the bottom of Figure 3, recovered grid poses are shown, as well as a circle fit to the positions of one grid point. Table 4 shows the RMS errors of circle fits (again, relative to scene size, and given in percent). We note that the non-central model provides a significantly better reconstruction than the central one for the spherical catadioptric camera, which thus confirms the above observation. For the fisheye, the non-central calibration also performs better, but not as significantly. As for the hyperbolic catadioptric camera, the central model gives a better reconstruction though. This can probably be explained as follows. Inspite potential imprecisions in the camera setup, the camera seems to be sufficiently close to a central one, so that the non-central model leads to overfitting. Consequently, although the bundle adjustment's residual is lower than for the central model (which always has to be the case), it gives "predictions" (here, pose or motion estimation) which are unreliable.

**Calibration of a multi-camera system.** A multi-camera network can be considered as a single generic imaging system. As shown in Figure 4 (left), we used a system of three (approximately pinhole) cameras to capture three images each of a calibration grid. We virtually concatenated the images from the individual cameras and computed all projection rays and the three grid poses in a single reference frame (see Figure 4 (right)), using the algorithm outlined in section 3.

In order to evaluate the calibration, we compared results with those obtained by plane-based calibration (Sturm & Maybank, 1999; Zhang, 2000), that used the knowledge that the three cameras are pinholes. In both, our multi-camera calibration, and plane-based calibration, the first grid was used to fix the global coordinate system. We can thus compare the estimated poses of the other two grids for the two methods. This is done for both, the rotational and translational parts of the pose. As for rotation, we measure the angle (in radians) of the relative rotation between the rotation matrices given by the two methods, see columns $R_i$ in Table 5). As for translation, we measure the distance between the estimated 3D positions of the grids' centers of gravity (columns $t_i$ in Table 5) expressed in percent, relative to the scene size. Here, plane-based calibration is done separately for each camera, leading to the three rows of Table 5.

From the non-central multi-camera calibration, we also estimate the positions of the three optical centers, by clustering the projection rays and computing least squares point fits to them. The column "Center" of Table 5 shows the distances between optical centers (expressed in percent and relative to the scene size) computed using this approach and plane-based calibration. The discrepancies are low, suggesting that the non-central calibration of a multi-camera setup is indeed feasible.

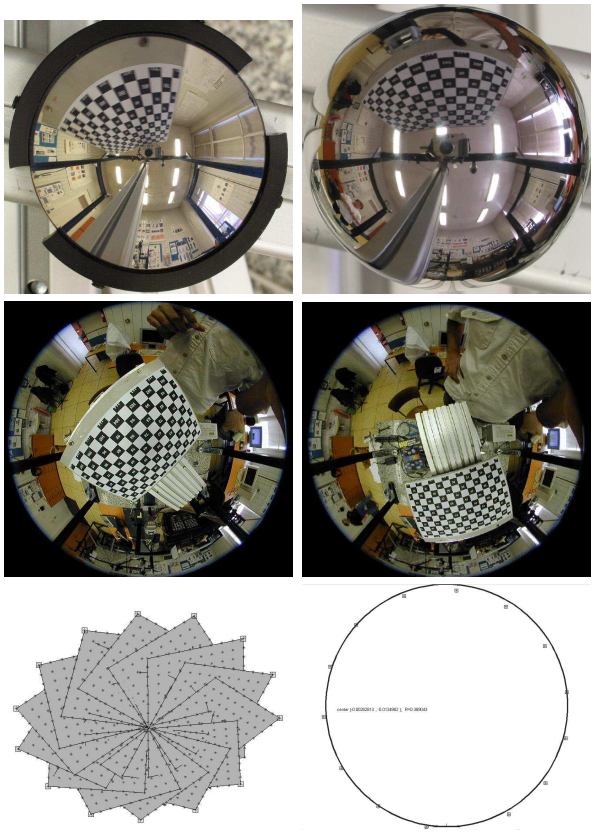

Figure 3: Top: sample images for hyperbolic and spherical catadioptric cameras. Middle: two images taken with a fisheye. Bottom: pose of calibration grids used to calibrate the fisheye (left) and a least squares circle fit to the estimated positions of one grid point (right).

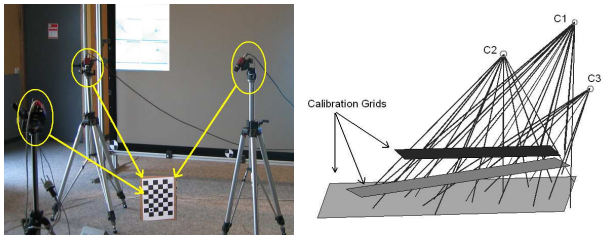


Figure 4: Multi-camera setup consisting of 3 cameras (left). Recovered projection rays and grid poses (right).

| Camera | $R_2$ | $R_3$ | $t_2$ | $t_3$ | Center |
|---|---|---|---|---|---|
| 1 | 0.0117 | 0.0359 | 0.56 | 3.04 | 2.78 |
| 2 | 0.0149 | 0.0085 | 0.44 | 2.80 | 2.17 |
| 3 | 0.0088 | 0.0249 | 0.53 | 2.59 | 1.16 |

Table 5: Evaluation of non-central multi-camera calibration relative to plane-based calibration. See text for more details.

**Structure-from-motion with hybrid camera setups.** We created hybrid camera setups by taking images with both, pinhole and fisheye cameras. Each camera was first calibrated individually using our approach of section 3. We then estimated the relative pose of two cameras (or, motion), using the approach

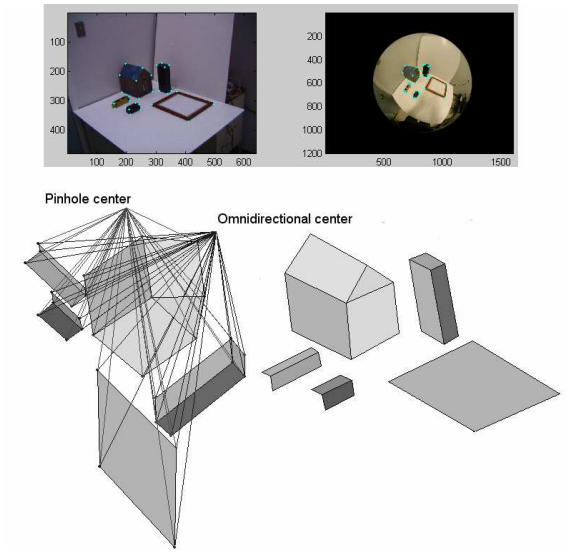


Figure 5: Combination of a pinhole and a fisheye camera. Top: input images and matching points. Bottom: estimated relative pose and 3D model.

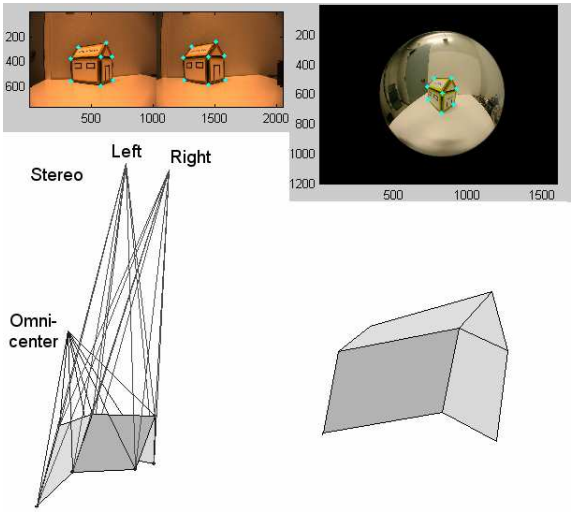


Figure 6: Combination of a stereo system and a fisheye camera. Top: input images and matching points. Bottom: estimated relative pose and 3D model.

outlined in section 4.2 and manually defined matches. Then, 3D structure was computed by reconstructing 3D points associated with the given matches.

Figure 5 shows this for a combination of a pinhole and a fisheye camera, and figure 6 for a combination of a stereo system and a fisheye. Here, the stereo system is handled as a single, non-central camera. Note that the same scene point usually appears more than once in the stereo camera. Therefore in the ray-intersection approach of section 5, we intersect three rays to find one 3D point here.

These results are preliminary: at the time we obtained them, we had not developed our full calibration approach of section 3.2, hence only the central region of the fisheye camera was calibrated and used. Nevertheless, the qualitatively correct results demonstrate that our generic structure-from-motion algorithms work, and actually are applicable to different cameras, or combinations thereof.

## 8 CONCLUSIONS

We have reviewed calibration and structure-from-motion tasks for the general non-central camera model. We also proposed a multi-view geometry for non-central cameras. A natural hierarchy of camera models has been introduced, grouping cameras into classes depending on, loosely speaking, the spatial distribution of their projection rays. We hope that the theoretical work presented here allows to define some common ground for recent efforts in characterizing the geometry of non-classical cameras.

The feasibility of our generic calibration and structure-from-motion approaches has been demonstrated on several examples. Of course, more investigations are required to evaluate the potential of these methods and the underlying models.

Among ongoing and future works, there is the adaptation of our calibration approach to axial and other camera models as well as first ideas on self-calibration for the general imaging model. We also continue our work on bundle adjustment for the general imaging model, cf. (Ramalingam et al. 2004), and the exploration of hybrid systems, combining cameras of different types (Sturm, 2002; Ramalingam et al. 2004).

**Acknowledgements.** This work was partially supported by the NSF grant ACI-0222900 and by the Multidisciplinary Research Initiative (MURI) grant by Army Research Office under contract DAA19-00-1-0352.

## REFERENCES

**References from Journals:**

Baker, S. and Nayar, S.K., 1999. A Theory of Single-Viewpoint Catadioptric Image Formation. *IJCV*, 35(2), pp. 1-22.

Chen, C.-S. and Chang, W.-Y., 2004. On Pose Recovery for Generalized Visual Sensors. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 26(7), pp. 848-861.

Geyer, C. and Daniilidis, K., 2002. Paracatadioptric camera calibration. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24(5), pp. 687-695.

Haralick, R.M., Lee, C.N., Ottenberg, K. and Nolle, M., 1994. Review and analysis of solutions of the three point perspective pose estimation problem. *International Journal of Computer Vision*, 13(3), pp. 331-356.

Hartley, R.I. and Sturm, P., 1997. Triangulation. *Computer Vision and Image Understanding*, 68(2), pp. 146-157.

Longuet-Higgins, H.C., 1981. A Computer Program for Reconstructing a Scene from Two Projections. *Nature*, 293, pp. 133-135.

Pajdla, T., 2002b. Stereo with oblique cameras. *International Journal of Computer Vision*, 47(1), pp. 161-170.

Peleg, S., Ben-Ezra, M. and Pritch, Y., 2001. OmniStereo: Panoramic Stereo Imaging. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 23(3), pp. 279-290.

Zhang, Z., 2000. A flexible new technique for camera calibration. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(11), pp. 1330-1334.

Zomet, A., Feldman, D., Peleg, S. and Weinshall, D., 2003. Mosaicing New Views: The Crossed-Slit Projection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 25(6), pp. 741-754.

**References from Books:**

Gruen, A. and Huang, T.S. (editors), 2001. *Calibration and Orientation of Cameras in Computer Vision*, Springer-Verlag.

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

**References from Other Literature:**


Non-central cameras for 3D reconstruction. Technical Report CTU-CMP-2001-21, Center for Machine Perception, Czech Technical University, Prague.

Bakstein, H. and Pajdla, T., 2001. An overview of non-central cameras. *Computer Vision Winter Workshop*, Ljubljana, Slovenia, pp. 223-233.

Barreto, J. and Araujo, H., 2003. Paracatadioptric Camera Calibration Using Lines. *International Conference on Computer Vision*, Nice France, pp. 1359-1365.

Champleboux, G., Lavallée, S., Sautot, P. and Cinquin, P., 1992. Accurate Calibration of Cameras and Range Imaging Sensors: the NPBS Method. *International Conference on Robotics and Automation*, Nice, France, pp. 1552-1558.

Faugeras, O. and Mourrain, B., 1995. On the Geometry and Algebra of the Point and Line Correspondences Between $N$ Images. *International Conference on Computer Vision*, Cambridge, MA, USA, pp. 951-956.

Geyer, C. and Daniilidis, K., 2000. A unifying theory of central panoramic systems and practical applications. *European Conference on Computer Vision*, Dublin, Ireland, Vol. II, pp. 445-461.

Gremban, K.D., Thorpe, C.E. and Kanade, T., 1988. Geometric Camera Calibration using Systems of Linear Equations. *International Conference on Robotics and Automation*, Philadelphia, USA, pp. 562-567.

Grossberg, M.D. and Nayar, S.K., 2001. A general imaging model and a method for finding its parameters. *International Conference on Computer Vision*, Vancouver, Canada, Vol. 2, pp. 108-115.

Hartley, R.I. and Gupta, R., 1994. Linear Pushbroom Cameras. *European Conference on Computer Vision*, Stockholm, Sweden, pp. 555-566.

Hicks, R.A. and Bajcsy, R., 2000. Catadioptric Sensors that Approximate Wide-angle Perspective Projections. *Conference on Computer Vision and Pattern Recognition*, Hilton Head Island, USA, pp. 545-551.

Neumann, J., Fermüller, C. and Aloimonos, Y., 2003. Polydioptric Camera Design and 3D Motion Estimation. *Conference on Computer Vision and Pattern Recognition*, Madison, WI, USA, Vol. II, pp. 294-301.

Nistér, D., 2003. An Efficient Solution to the Five-Point Relative Pose Problem. *Conference on Computer Vision and Pattern Recognition*, Madison, WI, USA, Vol. II, pp. 195-202.

Nistér, D., 2004. A Minimal Solution to the Generalized 3-Point Pose Problem. *Conference on Computer Vision and Pattern Recognition*, Washington DC, USA, Vol. 1, pp. 560-567.

Pajdla, T., 2002a. Geometry of Two-Slit Camera. Technical Report CTU-CMP-2002-02, Center for Machine Perception, Czech Technical University, Prague.

Pless, R., 2003. Using Many Cameras as One. *Conference on Computer Vision and Pattern Recognition*, Madison, WI, USA, Vol. II, pp. 587-593.

Ramalingam, S., Lodha, S. and Sturm, P., 2004. A Generic Structure-from-Motion Algorithm for Cross-Camera Scenarios. *5th Workshop on Omnidirectional Vision, Camera Networks and Non-Classical Cameras*, Prague, Czech Republic, pp. 175-186.

Shum, H.-Y., Kalai, A. and Seitz, S.M., 1999. Omnivergent Stereo. *International Conference on Computer Vision*, Kerkyra, Greece, pp. 22-29.

Sturm, P., 2002. Mixing catadioptric and perspective cameras. *Workshop on Omnidirectional Vision*, Copenhagen, Denmark, pp. 60-67.

Sturm, P. and Maybank, S., 1999. On Plane-Based Camera Calibration: A General Algorithm, Singularities, Applications. *Conference on Computer Vision and Pattern Recognition*, Fort Collins, CO, USA, pp. 432-437.

Sturm, P. and Ramalingam, S., 2003. A Generic Calibration Concept – Theory and Algorithms. Research Report 5058, INRIA.

Sturm, P. and Ramalingam, S., 2004. A generic concept for camera calibration. *European Conference on Computer Vision*, Prague, Czech Republic, pp. 1-13.

Swaminathan, R., Grossberg, M.D. and Nayar, S.K., 2003. A perspective on distortions. *Conference on Computer Vision and Pattern Recognition*, Madison, WI, USA, Vol. II, pp. 594-601.

Yu, J. and McMillan, L., 2004. General Linear Cameras. *European Conference on Computer Vision*, Prague, Czech Republic, pp. 14-27.