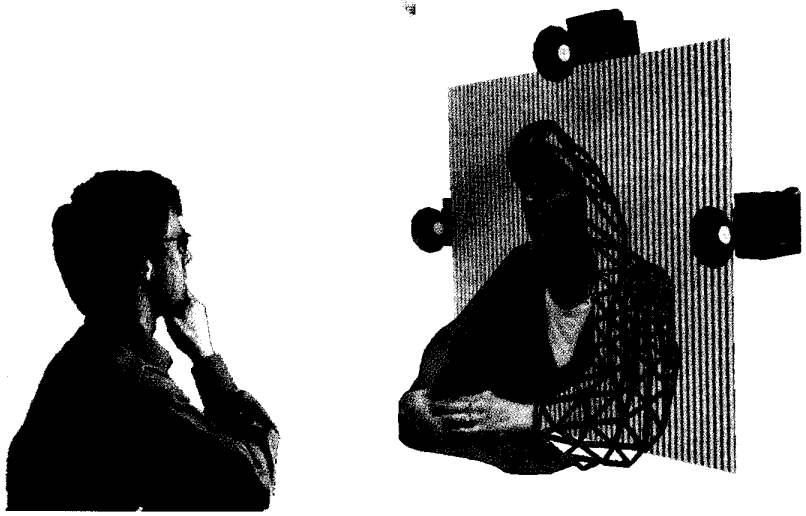
Multi-Camera Systems



Calibration and Applications

Federico Pedersini, Augusto Sarti, and Stefano Tubaro

Most of the methods for estimating the 3D structure of a scene through image analysis require an accurate a priori knowledge of the acquisition system's model. The parameters of this model can be estimated through a process called *camera calibration* [1], [2], [3], [4], which is based on the analysis of image features of one or more views. The targets that originate such features can be "artificial," i.e., *fiducial marks* that have been intentionally added to the scene, or "natural," i.e., natural object features such as vertices or corners. The estimation procedure varies depending on the structure and on the available a priori information. One common approach to camera parameter estimation is to use a rigid *target-set* that occupies part of the 3D viewing space, with a priori known geometrical characteristics.

When the 3D targets' coordinates are known beforehand, it is possible to use this information together with the image coordinates to estimate the camera parameters (*strong* calibration). Although the estimation accuracy is most influenced by that of the camera model [7], a major bottleneck is represented by the reliability of the 3D targets' coordinates. Due to the high cost of accurate measurement procedures, the only way to improve the performance of the parameter estimation process is to improve the available measurements' accuracy through a self-calibration [8] strategy.

The most extreme self-calibration case is given when a number of targets (artificial or natural) are scattered in the scene volume in unknown locations. Without a priori information on the targets, the increased number of unknowns makes this *blind calibration* problem undetermined [5], as it does not allow us to recover the whole geometry of the camera system [6]. The self-calibration problem, in fact, can become solvable when some a priori information on the targets or on the cameras is available. However, even when solvable, self-calibration is an ill-conditioned problem, therefore it is important to exploit all we know about the camera system and the scene. In fact, some rough information on the target-set (e.g. statistical information on the target's coordinates such as nominal position, and measurement's uncertainty) or on the cameras (e.g., focal length) is often available or can be readily measured. We will show that, if such information is fairly unbiased, it can be refined through self-calibration while estimating the parameters of the acquisition system.

One crucial problem of calibration strategies is their range of validity. The estimated parameters are, in fact, expected to hold accurate only within the 3D volume

1053-5888/99/$10.00©1999IEEE

"spanned" by the target-set [9]. As the target-set plays the role of *training set*, it should be designed in such a way to be "statistically representative" of the scene to be reconstructed. Accurate results are, in fact, obtained when the targets properly "fill up" the volume of interest, which means that the target-frame should be as large as the scene, with obvious difficulties in the calibration procedure. In order to overcome this difficulty, we can virtually "enlarge" a target-set of modest size through the acquisition of several of its views in different positions, so that the union of all targets will fill up volume of interest. Of course, every time we move the target-frame we introduce six new positional unknowns. Consequently, unless we are able to force the frame into pre-determined positions through some expensive high-precision positioning device, the only feasible solution is to embed the motion parameter estimation into the calibration process.

The temporal range of validity of calibration can also become a critical issue, especially when acquiring a long video sequence. Camera calibration is, in fact, very sensitive to mechanical shocks, vibrations and even thermal changes on both the cameras and the supports. This parameter drift can easily cause significant 3D reconstruction errors. One obvious way to overcome this problem is to use expensive, heavy and rigid camera supports. Our approach is, instead, to detect and track any changes in the acquisition system and correct the camera parameters "on the fly" using only the image coordinates of natural scene features.

The Camera Geometry

The camera model (see Fig. 1) that we adopted is basically an enhanced pinhole model that includes the nonlinear distortion of the optical lens and the offset between the principal point (point of orthogonal intersection between optical rays and image plane) and the center of the image. According to this model, the world coordinates $\mathbf{p}_w = [x_w \quad y_w \quad z_w]^T$ of a point of the 3D scene are mapped onto the (undistorted) image coordinates $\mathbf{p} = [x \quad y]^T$ through a relationship of the form

$$k\begin{bmatrix}\mathbf{p}\\ 1\end{bmatrix} = \mathbf{P}\begin{bmatrix}\mathbf{p}_w\\ 1\end{bmatrix}, \quad \mathbf{P} = \mathbf{K}[\mathbf{I}|\mathbf{0}]\mathbf{M} \tag{1}$$

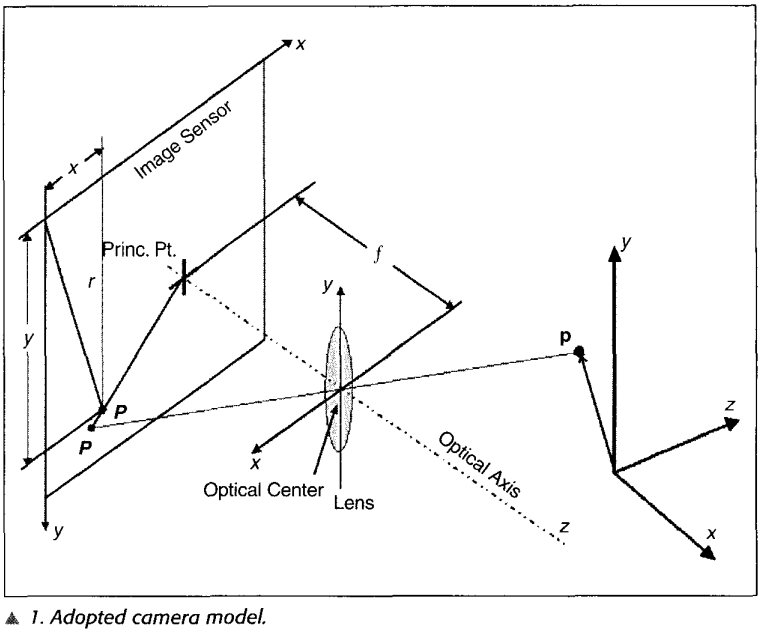
where k is a constant which normalizes to one the third element of the vector on the left hand side, $\mathbf{M}$ is the rigid motion matrix that maps the world-coordinates $\mathbf{p}_w$ of a point onto camera-coordinates $\mathbf{p}_c = [x_c \quad y_c \quad z_c]^T$, while $[\mathbf{I}|0]$ is the perspective projection matrix and $\mathbf{K}$ is a matrix that accounts for intrinsic camera parameters such as the focal distance f and the offset $\mathbf{p}_0 = [x_0 \quad y_0]^T$ of the principal point from the image center

$$\mathbf{K} = \begin{bmatrix}-f & 0 & x_0\\ 0 & -f & y_0\\ 0 & 0 & 1\end{bmatrix}, \quad [\mathbf{I}|\mathbf{0}] = \begin{bmatrix}1 & 0 & 0 & 0\\ 0 & 1 & 0 & 0\\ 0 & 0 & 1 & 0\end{bmatrix},$$
$$\mathbf{M} = \begin{bmatrix}\mathbf{R} & \mathbf{t}\\ \mathbf{0}^T & 1\end{bmatrix} \tag{2}$$

where $\mathbf{R}$ is the rotation matrix and $\mathbf{t}$ is the translation vector. The image coordinates can be easily expressed in pixel if the pixel size $\mathbf{d} = [d_x \quad d_y]^T$ is known.

Lens distortion, which is modeled as a nonlinear shift of the image points from their ideal perspective projection, is often well described by a polynomial model [10]. In order to accurately model lens distortion [11] both of its *radial* and *tangential* components should be considered [12], although in some cases the tangential component can be assumed as negligible [1]. In this last case the polynomial model becomes $r_u = r_d(1 + k_3 r_d^2 + k_5 r_d^4 = \ldots)$, in which the undistorted image coordinates $\mathbf{p}_u = [x_u \quad y_u]^T$ are written as a function of the distorted ones, r_d and r_u being the distances between the principal point and the distorted and undistorted image points, respectively. The first two coefficients of the power series k_3 and k_5 are often sufficient for an accurate parameterization of the radial distortion [7].

An alternative way of incorporating lens distortion into the camera model is to adopt a single (larger) projection matrix that maps a vector containing the object coordinates (and, in some cases, mixed products of coordinates up to a certain order) onto the final image coordinates [3], [11]. This choice, however, represents a



▲ *1. Adopted camera model.*

non-physical over-parametrization of the camera model [1], while we prefer to work with a non-redundant set of parameters that retain a physical meaning, in order to be able to exploit all the a priori information on the camera setup.

Inversion of the Camera Model

Let $\mathbf{p}_w(i).i = 1,\ldots,N$, be the world-coordinates of the i-th target and let $\mathbf{c}_j,\ j = 1,\ldots,M$ be the parameter vectors of the M cameras. The image coordinates $\mathbf{p}^{(j)} = [x^{(j)}(i)\ \ y^{(j)}(i)]^T$ of the i-th target, as seen from the j-th camera, can be the written as a function of both camera system parameters and target's coordinates

$$\mathbf{p}^{(j)}(i) = g(\mathbf{m}_{i,j}) \tag{3}$$

where $\mathbf{m}_{i,j} = [\mathbf{p}_w^T(i)\ \ \mathbf{c}_j^T(i)]^T$. This global equation can be thought of as a *direct* formulation of the camera modeling problem. Roughly speaking, self-calibration can be seen as the problem of *inverting* this direct formulation with respect to $\mathbf{m}_{i,j}$. When the 3D coordinates of the targets are known and accurate, they can be embedded in the direct model (strong calibration), which now becomes

$$\mathbf{p}^{(j)}(i) = g(\mathbf{p}_w(i), \mathbf{m}_j) = g^{(i)}(\mathbf{m}_j), \mathbf{m}_j = \mathbf{c}_j. \tag{4}$$

In order to comply with the terminology that is normally used for inverse problems [13], we will collect into a single vector $\mathbf{p}$ all the available target's image coordinates $\mathbf{p}^{(j)}(i)\ i = 1,\ldots,N,\ j = 1,\ldots,M$, while the 2D vector space $\mathcal{P}$ that can be spanned by $\mathbf{p}$ will be called *observation space*. Similarly, we will define a global parameter vector $\mathbf{m}$, which contains all the model vectors ($\mathbf{m}_{i,j}, i = 1,\ldots,N, j = 1,\ldots,M$, in the self-calibration case, or $\mathbf{m}_j,\ j = 1,\ldots,M$, in the strong calibration case) and spans the so-called *model space* M. In accordance to this terminology, $g(.)$ will be referred to as *direct model*.

From a practical standpoint, the strong/self-calibration process consists of exploiting a large number of constraints that cumulate in a space made of a large number of coordinates. The projection of a 3D point onto an image plane, in fact, gives rise to a pair of equations per image coordinate. It is customary (and advisable) to use a redundant number of fiducial points with respect to the number of unknowns, so that the model space will result as *over-constrained* [1]. Consequently, the determination of the model will have to be performed through a process of minimization of a measure of the error between the observed data $\mathbf{p}$ and the data computed through the model parameter vector $\mathbf{m}$ [4], [7], [8]. For example, adopting the MSE as a measure of this error, we will have to compute

$$\hat{\mathbf{m}} = \text{argmin}_{\mathbf{m}} ||\mathbf{p} - g(\mathbf{m})||^2. \tag{5}$$

This global optimization process, which is often referred to as *bundle adjustment*, is clearly non-linear and a variety of methods can be used to determine the solution $\hat{\mathbf{m}}$. The procedures that are commonly adopted for such non-linear problems are all iterative [7]; therefore an accurate initialization of the minimization process could become crucial for preventing the algorithm from being trapped into some local minima [1]. In order to take all the available information into account, each term of the cost function $||\mathbf{p} - g(\mathbf{m})||^2$ to be minimized can be weighed by a factor that takes into account the accuracy with which the 2D coordinates of the image point have been detected and the accuracy with which the coordinates of the corresponding 3D point are known [8], [14].

Some Remarks on Inverse Problems

Due to the limited image resolution and the unavoidable noise that affects the measuring process, the data vector $\tilde{\mathbf{p}}$ generally differs from the data vector $\mathbf{p}$ that we would predict if the CCD sensor was noiseless and had infinite resolution and if our camera model was infinitely accurate.

As $\tilde{\mathbf{p}}$ is measured through the analysis of the luminance profiles of the acquired views, it is usually affected by errors [15], [16]. In order account for the measurement's uncertainty, a conditional probability density function (PDF) of the form $f_{\tilde{\mathbf{P}}|\mathbf{P}}(\tilde{\mathbf{p}}|\mathbf{p})$ can be defined, where random vectors and their instances are denoted by uppercase and lowercase letters, respectively. The direct model's uncertainty is modeled by a "spread-function," which is a conditional p.d.f. of the form $f_{\tilde{\mathbf{P}}|\mathbf{M}}(\tilde{\mathbf{p}}|\mathbf{m}) = S(\mathbf{p} - \mathbf{m})$ that becomes an ideal impulse $\delta(.)$ when the model is perfect.

We should not forget that some a priori information on the model parameters is usually available or it is easy to retrieve. In fact, at least a rough idea of the relative world-coordinates of the targets is normally known, and sometimes the nominal focal length of the cameras (or at least a range of values) is available as well. This a priori information should not be ignored, therefore we need to incorporate it [13] in the calibration/self-calibration process through the definition of some proper probability density functions.

A statistical description of the acquisition system is given by the PDF $f_{\mathbf{P}|\mathbf{M}}(\mathbf{p}|\mathbf{m})$. In general, the solution of our inverse problem is the value of $\mathbf{m}$ that maximizes the a posteriori information on the model's parameters $f_{\mathbf{M}|\mathbf{P}}(\mathbf{m}|\mathbf{p})$, which can be derived from $f_{\mathbf{P}|\mathbf{M}}(\mathbf{p}|\mathbf{m})$. By doing so, we perform a maximum likelihood estimation (MLE) of the form

$$\mathbf{m}_{ML} = \max_{\mathbf{m}} ||f_{\mathbf{M}|\mathbf{P}}(\mathbf{m}|\mathbf{p})||^2. \tag{6}$$

Furthermore, when the sources of uncertainty that affect our inverse problem can be modeled by a zero-mean Gaussian PdF, it is also possible to predict the accuracy of the solution of the inverse problem in a rather general fashion. The a posteriori covariance can be estimated using a relationship of the form

$$\mathbf{C}_{\mathrm{M|P}} = (\mathbf{G}^T \mathbf{C}_{\mathrm{P}}^{-1} \mathbf{G} + \mathbf{C}_{\mathrm{M}}^{-1})^{-1} \quad , \quad \mathbf{G} = \left(\frac{\partial g}{\partial \mathbf{m}} \right)_{\mathbf{m}=\mathbf{m}_{ML}} \tag{7}$$

where $\mathbf{G}$ is the Jacobian of the forward model, which represents a linearization of $g(\mathbf{m})$ at $\mathbf{m}_{ML}$, $\mathbf{C}_M$ is the a priori covariance matrix of the model's parameter vector and $\mathbf{C}_P$ is the covariance matrix associated to both the "forward modeling uncertainty" and the "experimental uncertainty" (i.e. the statistical relationship between $\tilde{\mathbf{p}}$ and $\mathbf{p}$).

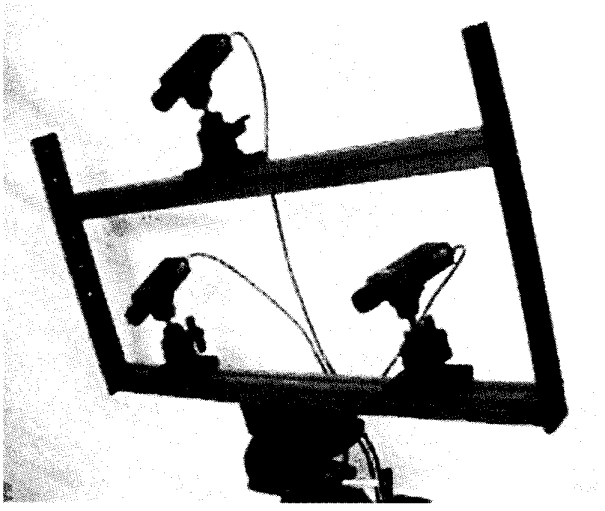
Notice that the a posteriori information on the model parameters ($\mathbf{C}_{\mathrm{M|P}}$) is obtained as a combination of a priori information ($\mathbf{C}_{\mathrm{M}}$) and information on the dispersion of the available data ($\mathbf{C}_{\mathrm{P}}$). The diagonal elements of $\mathbf{C}_{\mathrm{M|P}}$ represent the variance associated to the estimate of each model parameter $\mathbf{m}_{i,j}$. The other elements of $\mathbf{C}_{\mathrm{M}}$ can be used to estimate the correlation between the various parameters and to have an idea on how "separable" such parameters are [13]. In conclusion, an inverse problem can always be seen as a way of "translating" information from the data space P into the model space M, therefore the solution of a "well-posed" inverse problem should give an a posteriori uncertainty on the model parameters that is smaller than the a priori uncertainty [13].

Expanding the Range of Validity

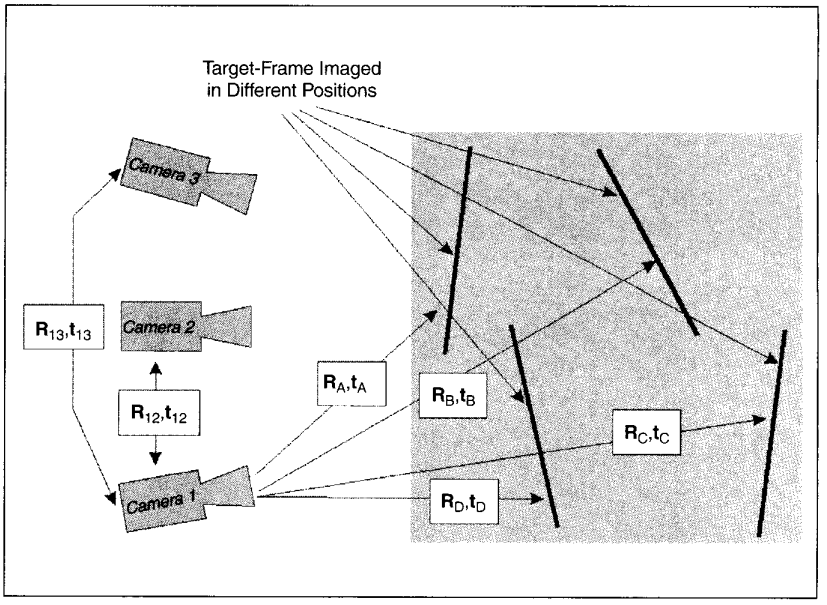
When the 3D coordinates of the fiducial points are known, we can adopt a strong calibration approach, although the (small) uncertainty on their positions can be included in the model through the computation of $f_{\mathrm{P|M}}(\mathbf{p}|\mathbf{m})$. In this case, the dimensionality of the model space is LM, where L is the number of parameters of the camera model and M is the number of cameras to be calibrated (an example of multi-camera rig is shown in Fig. 2). Indeed, in order for the strong calibration problem to be solvable, it is necessary to have at least as many constraints (equations) as unknowns. As each fiducial point generates a pair of equations per camera, a minimum of six independent (non-collinear) points is required for determining the parameters of the camera model.[2] However, since the problem is nonlinear and strongly ill-conditioned [12], a larger number of points should be considered.

Assuming that the pixel size is a known parameter[3], strong calibration can be performed with a planar target-set [1], which is much simpler to build and to measure than a 3D target-frame [17]. The main drawback of 2D target-sets is that they provide data that are quite correlated to each other. Furthermore, they occupy a rather limited portion of the scene, although this is true of all "portable" target-sets. It is well-known, in fact, that reliable results can only be achieved with a large number of targets that are well-distributed in the object space of interest [9]. In order to overcome such limitations, we virtually enlarge the target-set through the acquisition of several of its views. As the motion parameters of the target-set need to be embedded into the calibration process (see Fig. 3), this strong calibration process incorporates some self-calibration characteristics.

If we are considering V different positions of the target-set, then $6(V\text{-}1)$ new unknowns must be added. On the other hand, a total of $F = \sum_{k=1}^{V} 2H_{jk}$ equations can



▲ 2. Example of multi-camera acquisition system.



▲ 3. Schematic description of the multi-view, multi-camera approach to simple/self-calibration. A simple 2D target is used for calibrating an M-camera acquisition system. As the positions and the orientations of the calibration frame are a priori unknown, they need to be estimated through calibration.

be written, H_{jk} being the number of targets that are imaged in the k-th view taken from the j-th camera ($0 \le H_{jk} \le H$, H being the total number of targets of the calibration frame). In the previous section, we assumed that the acquisition system we intend to calibrate is multi-ocular. This assumption, however, is not restrictive. In principle, in fact, we could individually calibrate the cameras by following the above procedure. We should keep in mind, however, that a joint calibration of all cameras is generally more efficient and introduces a larger number of constraints in the parameter estimation process, with the result of reducing the risk of an erroneous estimation. As a matter of fact, if we consider that the motion of the target-frame from view to view is the same for all cameras, then each camera gives its contribution to the estimation of this motion. For this reason, the simultaneous calibration of all cameras of the acquisition system (MVMC approach) increases the well-posedness of the calibration problem, with the result of making the estimation easier and more reliable. With respect to the case in which one camera is calibrated with V views of the target-set, each additional camera adds L unknowns and approximately $2HV$ equations (assuming that all targets are imaged in the various acquisitions).

As in the strong calibration case, our self-calibration strategy is based on an MVMC approach. This way, we can virtually expand the target-frame and provide the estimator with more "independent" data. If an M-camera acquisition system is used for acquiring a set of V views per camera of a target-frame that contains H targets, then we have $2MVH > 3H + ML + 6(V-1)$. Notice that, on the left-hand side of this inequality is the number of constraints (two equations per viewed target per camera), under the simplifying assumption that the image coordinates of all targets can actually be determined. On its right-hand side is the number of unknowns to be estimated: $3H$ coordinates of the targets (such points are usually not exactly coplanar); ML camera parameters; and $6(V-1)$ parameters that characterize the motion of the target-frame. If, for example, we assume $L = 11$, the above inequality suggests us that a single-camera acquisition system would need at least $V = 2$ views for the self-calibration problem to be solvable. It is important to remember, however, that the self-calibration problem is generally undetermined, and is made solvable by the assumption that the errors on the 3D coordinates of the targets are limited (although not necessarily small) and have zero mean. In general, given the number of views, there is a minimum number H of targets below which the problem is undetermined. In practice, however, due to the ill conditioning of the problem, it is customary to use a number of targets that will make the problem largely over-determined.

Cruising the Parameter Space

As shown in the previous sections, calibration is an inverse problem to be approached through global nonlinear optimization. With this approach, which is commonly adopted in photogrammetry, the search space is made of a large number of unknowns (especially for self-calibration) and the cost function to be minimized is highly nonlinear. These facts make the search for the global minimum quite a difficult problem to solve. As a matter of fact, a number of solutions have been proposed in the literature, whose aim is to make the optimization problem solvable with a reasonable computational cost and with minimum risk of settling for a relative minimum. A simple solution is to use all the available a priori information on the camera parameters and on the target's coordinates. For example, if some rough information on the target's coordinates or the focal length is available, one could first determine a rough approximation of the other parameters by taking the available information for granted, and then refining all the parameters through a final bundle adjustment. As the first rough solution brings the algorithm closer to the global minimum, the final global optimization step can be speeded up and carried out more safely. When the search space is prohibitively large and the available a priori information is not sufficient to get close enough to the global minimum, or when we need a fast implementation of the estimation strategy, some additional constraints need be exploited.

Projective constraints and invariants are extensively used in computer vision for separating the estimation of intrinsic (physical) parameters from that of the extrinsic parameters (camera position and orientation), and for finding closed-form expressions to reduce the dimensionality of the search space. This approach consists of "peeling the layers" off a stratified model of vision, from projective to affine to euclidean. The first step is projective calibration, in which a camera projection matrix $\mathbf{P}$ [see (1)] is determined for each camera from a number of image correspondences [18], [19]. This operation is usually based on projective constraints that can be bilinear (epipolar constraint), trilinear (trifocal tensor), or multilinear, depending on the number of views that are being considered simultaneously [20].

According to the epipolar constraint, two stereo-corresponding optical rays are bound to be coplanar. This fact can be expressed in closed form as

$$(\mathbf{p}^{(2)})^T \mathbf{F} \mathbf{p}^{(1)} = 0\ ,\quad \mathbf{F} = (\mathbf{K}^{(2)})^{-T} \mathbf{R} \mathbf{S} (\mathbf{K}^{(1)})^{-1},$$
$$\mathbf{S} = \begin{bmatrix} 0 & -t_z & t_y \\ t_z & 0 & -t_x \\ -t_y & t_x & 0 \end{bmatrix} \tag{8}$$

where $\mathbf{p}^{(1)}$ and $\mathbf{p}^{(2)}$ are the image coordinates of two corresponding points as seen from the two cameras and $\mathbf{F}$ is the so-called fundamental matrix [18], which can be de-

termined in closed form if a number of corresponding image coordinates are available [19].

Notice that, from image correspondences alone, the camera projection matrix $\mathbf{P}$ can only be recovered up to a projective transformation. However, it is reasonable to assume that the intrinsic parameter $\mathbf{K}_i$ of (2) does not account for any skew. This restricts the ambiguity of the reconstruction to metric ($\mathbf{P}$ can be recovered up to a similarity transformation) [21], [6].

Given two views and the fundamental matrix that expresses the projective relationship between them, there are a variety of methods that allow us to determine the projection matrices (see, for example, [18] and [19]). Such methods usually exploit the fact that the projection matrices can be determined up to a similarity transformation so that the projection matrix of the first view can be simply chosen as $\mathbf{P}^{(1)} = [\mathbf{I} \mid \mathbf{0}]$. This choice results in a simplification of the procedure for determining the other projection matrix.

When more than two views are available, the determination of the projection matrix can be made more robust by adopting multilinear constraints. For example, one constraint that is often exploited when using three views is the trifocal tensor [20], which determines the position of a primitive in one image, given the position in other two.

Once the projection matrices $\mathbf{P}^{(k)}$ are available for all the views, the next step consists of determining the intrinsic and extrinsic matrices $\mathbf{K}^{(k)}$ and $\mathbf{M}^{(k)}$ that generated them. The literature is rich with methods for determining such matrices using additional geometric constraints. One solution often adopted consists of applying constraints on the intrinsic camera parameters through the absolute conic [22], [23]. The absolute conic is a set of points of imaginary projective coordinates $[x, y, z, t]^T$ that lie on the plane at infinity ($t = 0$) and satisfy the equation $x^2 + y^2 + z^2 = 0$. One remarkable property of the absolute conic is that it does not vary under scaled Euclidean transformations. Thus, its projection onto the image planes is invariant under rigid displacements of the camera (if the intrinsic parameters remain unchanged). The fact that the image of the absolute conic (IAC) depends only on the matrix $\mathbf{K}$ of the intrinsic camera parameters is confirmed by the fact that its equation is of the form $\mathbf{p}^T\mathbf{K}^{-T}\mathbf{K}^{-1}\mathbf{p} = 0$. As a consequence, if the matrix $\mathbf{K}^{-T}\mathbf{K}^{-1}$ of this conic can be found, we can determine $\mathbf{K}$ through Choleski factorization. In general, it is easier to work with the inverse KKT of the IAC matrix, which is called dual image of the absolute conic (DIAC).

The quadric of planes that are tangent to the absolute conic is called the absolute quadric, and is generally represented by a 4×4 symmetric rank-3 dual matrix Ω. If $\mathbf{T}$ transforms points from $\mathbf{p}_w$ to $\mathbf{T}\mathbf{p}_w$, then it transforms Ω onto $\mathbf{T}\Omega\mathbf{T}^T$. If $\mathbf{T}$ is a similarity transformation, then Ω remains unchanged. When the transformation is the projection matrix $\mathbf{P}$, then the absolute quadric is mapped onto the image of the absolute conic whose dual matrix is $\mathbf{P}\Omega\mathbf{P}^T$. A comparison with the DIAC yields the so-called Kruppa constraint

$$\lambda\mathbf{K}\mathbf{K}^T = \mathbf{P}\Omega\mathbf{P}^T \tag{9}$$

which relates the dual of the image of the absolute conic to the absolute quadric [24]. Since we know the projective matrix (up to a change of basis), if some information is already available on the intrinsic camera parameters, the Kruppa constraint can be used for recovering the Euclidean geometry. For example [23], [24], knowing the coordinates (x_0, y_0) of the principal point, it is possible to rewrite the Kruppa constraint as

$$\lambda_k \begin{bmatrix} f_k^2 & 0 & 0 \\ 0 & f_k^2 & 0 \\ 0 & 0 & 1 \end{bmatrix} = \mathbf{P}^{(k)} \begin{bmatrix} f_1^2 & 0 & 0 & a_1 f_1^2 \\ 0 & f_1^2 & 0 & a_2 f_1^2 \\ 0 & 0 & 1 & a_3 \\ a_1 f_1^2 & a_2 f_1^2 & a_3 & (a_1^2 f_1^2 + a_2^2 f_1^2 + a_3^2) \end{bmatrix} (\mathbf{P}^{(k)})^T \tag{10}$$

where k is the index that specifies the considered view, assuming a reference frame attached to the first camera so that $\mathbf{P}^{(1)} = [\mathbf{I} \mid \mathbf{0}]$. In fact, from (10) we recognize that Ω is symmetric and rank 3. The above version of the Kruppa constraint allows us to specify a total of $4(n-1)$ equations (four equations for each one of the n-1 views, as the first view is excluded) in the four unknowns f_1^2, a_1, a_2, and a_3. More specifically, if y_{ij} is the (i,j) element of the maitrix $\Gamma = \lambda_k \mathbf{K}_k \mathbf{K}_k^T$, then the four equations associated with the k-th view ($k = 2 \ldots n$) are $y_{11} = y_{22}$ and $y_{12} = y_{13} = y_{23} = 0$. As a consequence, we need at least two views in order to determine the focal lengths. Once Ω has been computed, and $\lambda_k \mathbf{K}_k \mathbf{K}_k^T$ is available, the above four equations allow us to determine $\mathbf{K}_k$. In order to do so, since Γ_k is already diagonal, it is not necessary to compute its Choleski factorization, as it suffices to let y_{33} be the scale factor λ_k. The focal length turns out to be

$$f_k = \sqrt{y_{11} / y_{33}}.$$

Now that the intrinsic matrices are known, the epipolar constraint between two views becomes

$$(\mathbf{q}^{(k)})^T \mathbf{E}\mathbf{q}^{(m)} = 0,$$

where

$\mathbf{q}^{(k)} = (\mathbf{K}^{(k)})^{-1}\mathbf{p}^{(k)}$, $\mathbf{q}^{(m)} = (\mathbf{K}^{(m)})^{-1}\mathbf{p}^{(m)}$ and $\mathbf{E} = \mathbf{R}\mathbf{S}$ is called *essential matrix*, which only contains extrinsic parameters. From the essential matrix it is possible to algebraically determine $\mathbf{R}$ and $\mathbf{t}/|\mathbf{t}|$ through a process based on singular value decomposition [19]. The scale factor of the translation can finally be (at least approximately) determined by exploiting the a priori knowledge on the actual size of the calibration target-set or the actual distance

In principle, the target-set's complexity could be minimized by reducing the number of targets to the minimum, while increasing the number of positions from which the target-frame is viewed.

between two of the scene features that originated the used image points.

This approach has the immediate advantage of dealing with the many variables in a more structured fashion, although it suffers from the drawback that it does not account for lens nonlinearities. Furthermore, the determination of the principal points is not carried out explicitly. In order to overcome this last problem, Borghese et. al. [25] estimate the principal points of a binocular camera system through a nonlinear minimization process based an evolutionary approach, while projective constraints and invariants are used for computing the rest of the parameters in closed form.

The above approach can be used for a careful initialization of a global optimization process, and to achieve results of considerable accuracy with limited risk of encountering relative minima and a heavy reduction of the computational load. However, we performed some estimation experiments in which the nonlinear distortion parameters were estimated with the principal points while the remaining parameters were computed explicitly. More specifically, we implemented a loop in which the optimization of distortion parameters and principal points is cascaded with the explicit computation of the remaining parameters until a stable configuration is reached [26]. The cost function is based on the agreement between the actual data and the data predicted with the estimated model, and on the degree of satisfaction of other constraints. As our camera system is trinocular, the system uses both bilinear and trilinear constraints.

The tests performed on this parameter estimation approach seem to confirm its effectiveness, as its adoption speeds up the computational time of almost two orders of magnitude. We will see in the next section that this approach has the advantage of increasing the number of constraints considerably, allowing us to further reduce the complexity of the target-set.

Implementation

When the target-frame is planar, it is very easy to roughly measure the relative targets' coordinates. In this case, the global minimum can be safely reached through an iterative process in which the number of unknowns is progressively increased at each step. For example [27], [28], in the case of strong calibration, we can first proceed with the individual calibration of each camera through a separate analysis of the V views of the target-frame. By averaging the V sets of camera parameters obtained from the individual calibrations, we obtain a good starting point for the next calibration step. This second step still concerns the individual cameras, but it uses all the available views simultaneously. At the end of this process, we obtain a refined version of the camera parameter vector $\mathbf{c}$ and M estimates of the vector $\mathbf{v}$ that describes the target positions. The last step consists of a global (all cameras) calibration based on the simultaneous analysis of all the views. This last optimization step refines the previous estimate of the parameters, which consists of the parameter vector $\mathbf{c}$ and an average between all the target-frame's motion vectors.

In the case of self-calibration, the camera parameters can be roughly estimated first through MVMC calibration by pretending that our a priori knowledge of the target's world coordinates is not affected by uncertainty. The whole vector $\mathbf{m}$ containing both targets' world coordinates and camera parameters can then be refined through a global minimization process.

Notice that even when the uncertainty of the targets' world coordinates is very limited, we can always estimate the uncertainty of the model's parameters through a linearization of the direct model. The uncertainty of the targets' world coordinates, in fact, can be treated in the same way as the uncertainty σ_{loc} of the image feature localization. For a first and rough estimate of this uncertainty, it is reasonable to assume that, if the targets are Gaussianly distributed over a planar frame around their nominal locations, their projections on the image plane remains Gaussianly distributed. Therefore, their standard deviation on the image plane is $\sigma_{\mathrm{P}} = \sigma_{\mathrm{P}_w} f / z_c$, where σ_{P_w} is the standard deviation of the uncertainty of the targets' world coordinates, f is the focal length of the camera, and z_c is the average distance between the camera and targets. Because the 3D measurements and the image feature localization error can be considered statistically independent, a measure of the uncertainty of the global data can be computed. The uncertainty can still be considered Gaussian, with variance $\sigma^2_{\tilde{\mathrm{P}}} = \sigma^2_{loc} + \sigma^2_{\mathrm{P}}$. The global covariance matrix $\mathbf{C}_{\tilde{\mathrm{P}}} = \sigma^2_{\tilde{\mathrm{P}}} \mathbf{I}_{2N}$ is then used to predict the uncertainty $\mathbf{C}_{\mathrm{M}}$ on the estimated model.

We have experimentally verified that the calibration's uncertainty (measured as the variance of the target's reconstruction error) is approximately proportional to the inverse of the number of targets (with a fixed number of target views). Let us assume, for example, that the acquisition system is made of three standard TV-resolution cameras (Fig. 2). Due to the geometry of the acquisition systems, and considering the fact that the orientation of the planar target needs to be reasonably tilted with respect to the camera's optical axes, it is usually reasonable to acquire four or five views of the target-frame. Assuming

that the accuracy of the feature localization is about 0.1 pixel, with a target-set placed at about 800 mm from the cameras (assumed as being about 500 mm apart from each other), we cannot expect the target's reconstruction accuracy to be any better than 0.1 mm. In such conditions, 15-20 targets turned out to be enough for maximum accuracy.

Minimizing the Target-Set

In principle, the target-set's complexity could be minimized by reducing the number of targets to the minimum, while increasing the number of positions from which the target-frame is viewed. For example, the estimation of a set of 11 parameters for each camera of a trinocular acquisition system requires a minimum of six targets (which can also be coplanar).[4] Below this minimum it is necessary to exploit additional projective constraints and invariants.

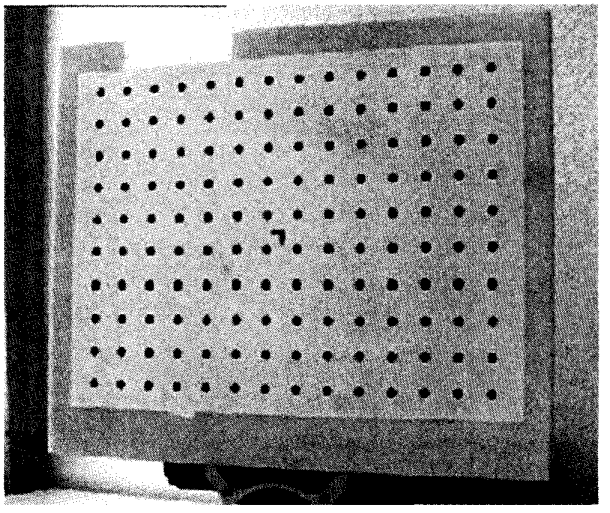
One interesting example of this strategy [25] was developed for a binocular acquisition system. A calibrated bar ending with two spherical targets at a known distance, is moved inside the working volume while the camera system acquires a sequence of images. As the target frame comprises just two targets, a fairly large number of images must be acquired before the desired accuracy in the estimated parameters is reached. The method becomes particularly convenient with stereo video sequences, which pose some limitations to the computational complexity of the approach. Instead of adopting a global minimization procedure, in fact, the method uses the projective constraints and invariants of the previous section in order to compute as many parameters as possible in closed form. The remaining parameters are estimated through a proper optimization process. More specifically, the focal lengths are determined using the properties of the absolute conic in the projective space while the extrinsic camera parameters are computed from the epipolar geometry up to a scale factor, which is determined from the actual length of the calibration bar. The principal points are estimated through a nonlinear minimization process, which is carried out through an evolutionary optimization approach. This method does not account for nonlinear lens distortion, however, we performed some experiments in which the distortion parameters were included in the nonlinear optimization step, confirming that the hypothesis of the absence of nonlinear distortion can be removed [26].

Although the above estimation method is not based on self-calibration, the need to measure the rigid bar beforehand with appropriate accuracy is not too strong a burden, as it consists of just a simple 1D measurement, whose accuracy will directly influence the quality of the camera parameter estimation.

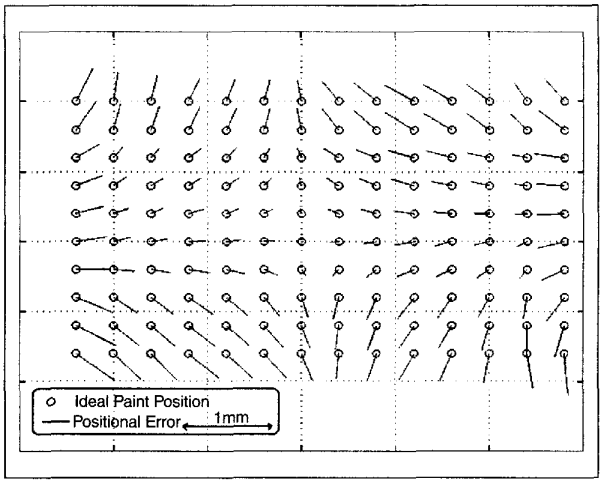
Making the Calibration Adaptive

In order to detect and track any camera parameter changes and correct them "on the fly" [29], [30], the image features (control points) to be analyzed need to be localized with high (sub-pixel) accuracy [31], [32]. If the control points are artificial targets, then they can be detected and localized through some advanced template matching process. In the case of natural object features, one solution is to search for corners or vertices (viewer-independent crossings between 3D edges) [33], [34], [35].

Vertices are characterized by the fact that the Laplacian of their luminance profile is zero. Furthermore, the Baudet operator $DET = I_{xx} I_{yy} - I_{xy}^2$ has a relative maximum (in all directions) in the proximity of vertices and, when applied to a set of progressively more filtered versions of the image, the maxima can be shown to be collinear. These two constraints can be used jointly for deter-



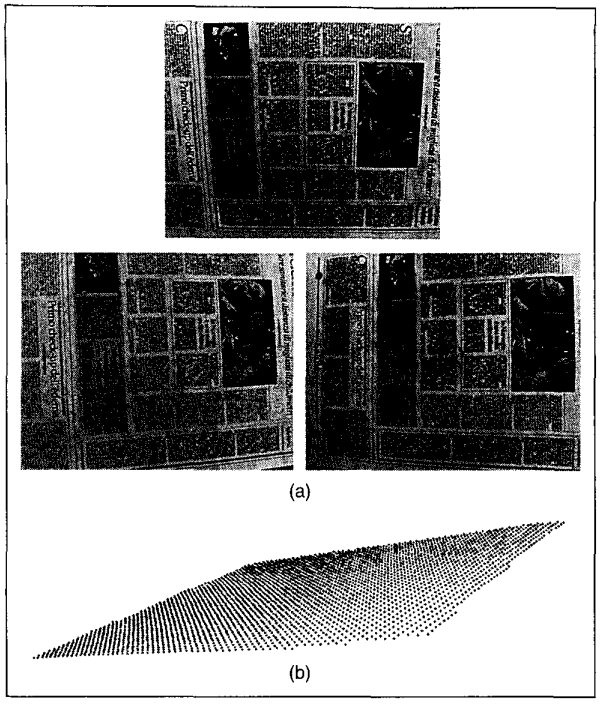
▲ *4. Example of low-cost target-set (laser-printed sheet of paper glued to a planar glass surface).*



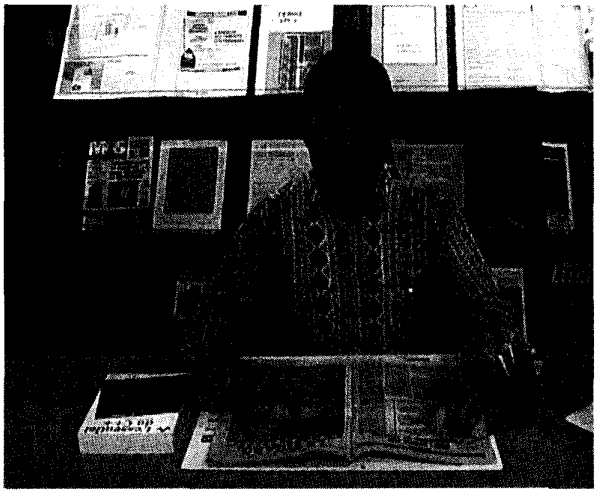
▲ *5. Targets' positional errors caused by the dragging action of the laser printer.*

mining a vertex with super-resolution accuracy. In order to do so, we can look for the zero-crossing of the Laplacian along the line of the maxima of the DET. The achieved results show that such improvements allow us to reach a localization accuracy that is better than 0.2 pixels. As far as feature matching is concerned, we use an M-partite matching algorithm based on the similarity of the local luminance profiles, which also takes the epipolar geometry specified by the current calibration into account. The matching process generates a set of M-tuples (M being the number of cameras) of homologous points, which can be back-projected onto the 3D scene space. Through a proper analysis of the magnitude and the temporal behavior of the back-projection error, we can reveal and characterize any incidental modification of the camera parameters. In order to correct the parameters on the fly, we keep track of the stable scene points and use them for re-calibration or self-calibration.

In order to validate the adaptive calibration technique, we conducted a series of experiments. In some cases, we simulated a change in the acquisition setup by modifying the calibration parameters. In others, we physically modified the camera setup (e.g., relative camera position or focal length). In all such cases, the system revealed a significant increment of the accuracy index, and, since the number of matched points was not significantly affected by the parameter change, we could use most of the matched points as control points and correct the calibration parameters. The accuracy of the corrected parameters turned out to be comparable with that of the original calibration in both cases of re-calibration and self-calibration.



▲ *6. Example of 3D reconstruction: triplet of original views of a flat newspaper's page (a) and oblique view of a set of 3D points of the newspaper's page, retrieved through area matching (b). The resulting standard deviation from planarity is 0.1 mm.*

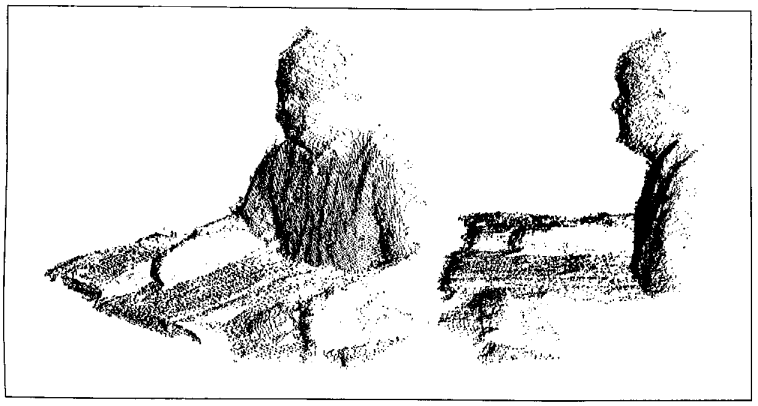


▲ *7. Example of teleconferencing scene.*

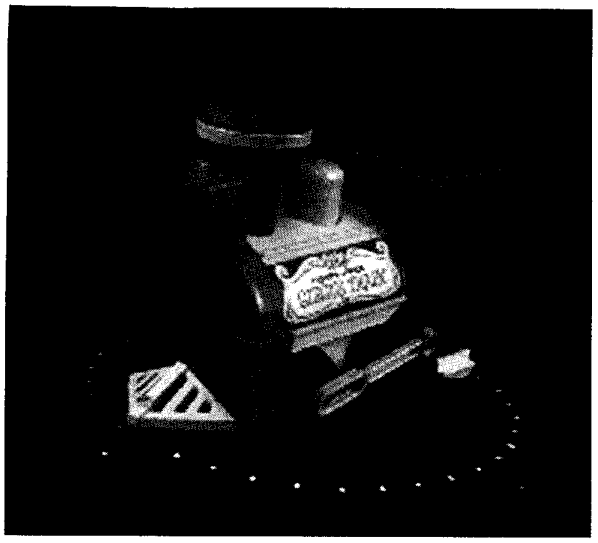
Experimental Results

In order to test the reliability of our calibration methods, we performed a series of tests in a variety of experimental conditions, using three standard TV-resolution CCD cameras (Fig. 1). A first series of tests was conducted on a high-quality target-frame whose targets, nominally scattered on a regular grid, were accurately measured through a photogrammetric procedure. The surface of the target-frame was a lightweight honeycomb-structured aluminum "wafer" for improved rigidity. MCMV self-calibration was performed using only the nominal target's coordinates, producing results whose accuracy was comparable with that obtained through strong calibration using the accurate measurements on the target's coordinates.

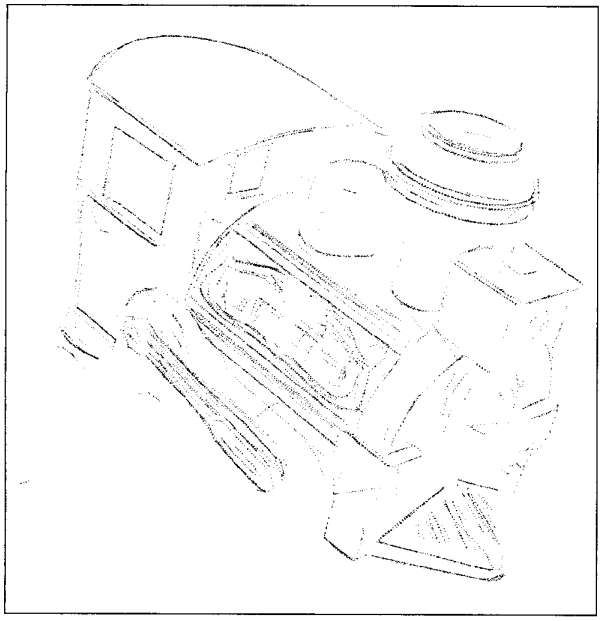
Some other self-calibration MCMV experiments were conducted with an inexpensive target frame (see Fig. 4) obtained by gluing an A4-sized sheet of laser-printed paper to a flat surface. In Fig. 5, the a priori location of the target points (defined on the printed pattern) and the corresponding positional error (estimated through self-calibration) are visualized. As we can see, the paper dragging action of the laser printer causes a positional error, as visually confirmed using a high-precision ruler. In conclusion, laser printers are able to guarantee high resolution, but poor positional accuracy; making this type of target-set suitable for self-calibration and not for strong calibration. However, as the positional error caused by the dragging action is not guaranteed to be unbiased, it is preferable to use a professional ink-jet plotter, as confirmed experimentally. In this case, in fact, we found a good agreement between a priori and a posteriori world coordinates of the targets; which confirms that such target-sets can be effectively used for strong calibration.



▲ 8. Two views of the 3D point clouds estimated through area matching.



▲ 9. Example of complex object.



▲ 10. Cloud of 3D points reconstructed from several triplets of views after fusion.

Finally, we finally carried out some experiments for evaluating the maximum accuracy that can be reached by a 3D reconstruction procedure based on stereo-correspondences [37], when using the above-described trinocular camera system, calibrated with our MCMV method. To do so, we considered a set of views of the target-frame that had not already been used for calibration. We estimated the distance between fiducial points through back-projection of their image-coordinates. The obtained accuracy was better than 0.2 mm, with an average distance of 2000 mm between cameras and object, and a maximum object size of approximately 1500 mm (corresponding to a relative accuracy of about 130 ppm). Similar results were found with some other 3D reconstruction experiments performed on a variety of test objects. In Fig. 6, an example of reconstruction of a flat surface (a newspaper page glued to a flat surface) from one triplet of views is reported. In this case, the resulting standard deviation from planarity was about 0.1 mm. In Figs. 7 and 8, a complex teleconferencing scene and its reconstruction from one triplet of views are shown. Finally, an example of reconstruction from several triplets of views is shown in Figs. 9 and 10. Data fusion is achieved by first estimating the egomotion from 3D curve matching, and by then merging the partial reconstructions.

Conclusions

In this article, we presented some simple and effective techniques for accurately calibrating a multi-camera acquisition system. The proposed methods were proven to be capable of accurate results even when using very simple calibration target-sets and low-cost imaging devices, such as standard TV-resolution cameras connected to commercial frame-grabbers. In fact, the performance of our calibration approach yielded results that were about the same as that of other traditional calibration methods based on large 3D target sets [1], [12]. The proposed calibration strategy is based on a multi-view, multi-camera approach. This was based on the analysis of a number of views of a simple calibration target-set placed in different (unknown) positions. Furthermore, the method is based on a self-calibration approach, which can refine the a priori knowledge of the world coordinates of the targets (even when such information is very poor) while estimating the parameters of the camera model. Finally, we proposed a method to make the calibration technique adaptive through the analysis of natural scene features, allowing the camera parameters to hold accurate throughout the acquisition session in the presence of parameter drift.

Acknowledgments

The authors wish to thank Dr. Alberto Borghese for the fruitful discussions about target-set minimization, and the useful suggestions on its implementation.

References

[1] R.Y. Tsai, "A versatile camera calibration technique for high-accuracy 3D machine vision metrology using off-the-shelf TV cameras and lenses," *IEEE J. on Robotics and Automation*, vol. RA-3, no. 4, pp. 323-344, Aug. 1987.

[2] Y.I. Aziz and H.M. Karara, "Direct linear transformation into object space coordinates in close-range photogrammetry," *Proc. Symp. Close-Range Photogrammetry*, Urbana, IL, pp. 1-18, 1971.

[3] Y. Yakimowsky and R. Cunningham, "A system for extracting three-dimensional measurements from a stereo pair of TV cameras," *Computer Graphics and Image Processing*, no. 7, pp. 195-210, 1978.

[4] H.A. Beyer, "Some aspects of the geometric calibration of CCD cameras," *ISPRS Intercomm. Conf. on Fast Processing of Photogrammetric Data*, Interlaken, 1987.

[5] O. Faugeras, "Stratification of three-dimensional vision: projective, affine, and metric representations," *J. Opt. Soc. of Am.*, vol. 12, no. 3, pp. 465-84, Mar. 1995.

[6] S. Bougnoux, "From projective to Euclidean space under any practical situation, a criticism of self-calibration," *IEEE 6th Intl. Conf. on Computer Vision*, Bombay, India, Jan. 1998, pp. 790-796.

[7] J. Weng, P. Cohen, and M. Herniou, "Camera calibration with distortion model and accuracy evaluation," *IEEE Trans. PAMI*, vol. 14, no. 10, pp. 965-980, Oct. 1992.

[8] W. Faig, *Manual of Photogrammetry*, 4th ed., American Society of Photogrammetry, 1990.

[9] G. Ferrigno, N.A. Borghese, and A. Pedotti, "Pattern recognition in 3D automatic human motion analysis," *ISPRS J. of Photogrammetry and Remote Sensing*, no. 45, pp. 227-246, 1990.

[10] P.R. Wolf, *Elements of Photogrammetry*, New York: McGraw-Hill, 1993.

[11] R. Lenz and U. Lenz, "New developments in high-resolution image acquisition with CCD area sensors," in *Optical 3-D Measurement Techniques II*, Gruen/Kahmen Eds., Wichmann, 1993.

[12] G.Q. Wei and S. De Ma, "Implicit and explicit camera calibration: Theory and Experiments," *IEEE Trans. PAMI*, vol. 16, no. 5, pp. 469-480, May 1994.

[13] A. Tarantola, *Inverse Problem Theory*, Amsterdam: Elsevier, 1987.

[14] D. Zhang, Y. Nomura, S. Fujii, "Error analysis and optimization of camera calibration," *Proc. of IEEE/RSJ Intl. Workshop on Intelligent Robots and Systems IROS-91*, Osaka, Japan, Nov. 3-5, 1991, pp. 292-296.

[15] D. Barbe, "Imaging Devices Using the Charge-Coupled Concept," *IEEE Proceedings*, vol. 63, no. 1, Jan. 1975.

[16] H.A. Beyer, "Geometric and radiometric analysis of a CCD-camera-based photogrammetric close-range system," Ph.D. Thesis, no. 51, Institut für Geodäsie und Photogrammetrie, ETH, Zürich, May 1992.

[17] C.F. Laizet, "Determination of video camera parameters in stereoscopic mode," *4th European Workshop on Three-Dimensional Television*, Oct. 20-21, 1993, Rome, Italy.

[18] O.D. Faugeras, "What can be seen in three dimensions with an uncalibrated stereo rig?" *ECCV '92, Lecture Notes in Computer Science*, Santa Margherita Ligure, Italy, pp. 563-578, May 1992.

[19] R.I. Hartley, "Estimation of relative camera positions for uncalibrated cameras," *ECCV '92. 2nd European Conf. on Comp. Vision*, Santa Margherita Ligure, Italy, May 1992, p. 579-87.

[20] P. Beardsley, P. Torr, and A. Zisserman, "3D model acquisition from extended image sequences," *4th European Conf. on Comp. Vision. ECCV '96*, Cambridge, UK, April 1996, pp. 683-95.

[21] M.E. Spetsakis and J. Aloimonos, "A unified theory of structure from motion," *Proc. DARPA IU Workshop*, pp. 271-283, 1990.

[22] B. Triggs, "Autocalibration and the absolute quadric," *IEEE Conf. on Computer Vision and Pattern Recognition*, San Juan, Puerto Rico, June 1997, pp. 609-614.

[23] M. Pollefeys, L. Van Gool, and M. Proesmans, "Euclidean 3D reconstruction from image sequences with variable focal lengths," *ECCV '96*, Cambridge, UK, April 1996, pp. 31-42.

[24] M. Pollefeys and L. Van Gool: "Self-calibration from the absolute conic on the plane at infinity." *7th Intl. Conf. On Computer Analysis of Images and Patterns*, Kiel, Germany, Sept. 10-12, 1997, pp. 175-82.

[25] N.A. Borghese and P. Cerveri, "Calibrating a video camera pair with a rigid bar," to appear in *Pattern Recognition*.

[26] P. Milani, F. Pedersini, A. Sarti, and S. Tubaro, "Self-calibration with a minimal target-set," ISPG Report no. 99.2.1. Politecnico di Milano, 1999.

[27] F. Pedersini, A. Sarti, and S. Tubaro, "Accurate and low-cost calibration and self-calibration of multi-camera acquisition systems," to appear in *EURASIP Signal Processing*.

[28] F. Pedersini, D. Pele, A. Sarti, and S. Tubaro, "Calibration and self-calibration of multi-ocular camera systems," *Intl. Workshop on Synthetic-Natural Hybrid Coding and Three-Dimensional (3D) Imaging*, Rhodes, Greece, Sept. 1997, pp. 81-84.

[29] F. Pedersini, A. Sarti, and S. Tubaro, "Tracking camera calibration in multi-camera sequences through automatic feature detection and matching," *IX EUSIPCO*, Rhodes, Greece, Sept. 1998, pp. 1081-1084.

[30] F. Pedersini, A. Sarti, and S. Tubaro, "Accurate feature detection and matching for the tracking of calibration parameters in multi-camera acquisition systems," *ICIP-98*, Chicago, IL, Oct. 1998, pp. 598-602.

[31] M. Armstrong, A. Zisserman, and R. Hartley, "Self-calibration from image triplets," *ECCV '96*, Cambridge, U.K., pp. 3-16.

[32] Z. Zhang, R. Deriche, O. Faugeras, and Q.T. Luong, "A robust technique for matching two uncalibrated images through the recovery of the unknown epipolar geometry," INRIA, Report no. RR 2273, 1994.

[33] L. Kitchen and A. Rosenfeld, "Gray-level corner detection," *Pattern Recognition Letters*, no. 1, 1982, pp. 95-102.

[34] G. Giraudon and R. Deriche, "On corner and vertex detection," *Intl. Conference on Computer Vision and Pattern Recognition*, Maui, Hawaii, June 1991, pp. 650-655.

[35] K. Rohr, "Recognizing corners by fitting parametric models," *Intl. J. of Computer Vision*, vol. 9, no. 3, 1992, pp. 213-230.

[36] F. Pedersini, A. Sarti, and S. Tubaro, "Accurate and simple geometric calibration of multi-camera systems," *EURASIP Signal Processing*, vol. 77 no. 3, 1999.

[37] P. Pigazzini, F. Pedersini, A. Sarti, and S. Tubaro, "3D area matching with arbitrary multiview geometry," To appear in *EURASIP Signal Processing: Image Communications*.

[38] F. Pedersini, A. Sarti, and S. Tubaro, "Egomotion eEstimation of a multicamera system through line correspondence," *ICIP-97*, Oct. 26-29, 1997, Santa Barbara, CA, pp. 175-178.

Endnotes

[1]Work supported in part by the ACTS Project "PANORAMA," Proj. No. AC092.

[2]It can be shown [1] that, when dealing with a single-camera system, all points should not be co-planar unless the pixel size is known beforehand.

[3]This is a reasonable assumption with digital cameras; or with CCD analog cameras with known ratio between pixel-clock and frame grabber's sampling rate.

[4]Six extrinsic parameters (three Euler angles and three translational components) plus five intrinsic parameters (principal point's offset, two coefficients of the radial distortion and the focal length).