# Optimization methods to calibrate a stereo rig with increased accuracy for vehicular applications

András Bódis-Szomorú, Tamás Dabóczi
Department of Measurement and Information Systems
Budapest University of Technology and Economics, Budapest, Hungary
Email: {bodis,daboczi}@mit.bme.hu, http://www.mit.bme.hu/∼bodis

***Abstract*—In this paper, several methods are presented to fully calibrate a stereo vision system for far-range outdoor applications. Traditionally, intrinsic camera parameters are determined from a high number of feature points, while pose estimation is solved with a few (tens). Close-range setups are not well suited for far-range applications, while far-range setups can easily deviate from planar and small measurement errors may result in large errors in the marker positions. Our Maximum Likelihood (ML) formulation to the stereo pose problem takes such inaccuracies into consideration. Herein, our earlier results are extended by decoupling pose estimation into an inter-camera and a rig-to-world pose problem to avoid relying on a small number of features with all extrinsic parameters. A high number of point-matches are extracted from image pairs acquired during on-line operation, and are incorporated into the procedure for off-line inter-camera pose estimation via the essential matrix. Using real and synthetic data, it is shown how the far-range arrangement and the matches can be used to adjust the camera models, even after a sound intrinsic calibration and ML pose estimation. The proposed methods rely on point matches between views and, thus, are expected to produce better reference for the evaluation of autocalibration methods that typically begin with establishing such correspondences.**

## I. Introduction

Vision-based 3D environment perception has become increasingly important in transportation in the last decade [1], [2]. Advanced Driver Assistance Systems aim to enhance safety and efficiency of transportation, and rely on on-board vision sensors that are available at low-cost for the automotive industry (e.g. compared to 3D LiDARs). Stereo vision solutions are popular for an instantaneous capturing of 3D information [3], [4]. Stereo vision eliminates the necessity of relying on non-generic assumptions about the observed scene, e.g. rigid vehicle suspension, flat road or constant lane width. However, the quality of 3D reconstruction is greatly influenced by the precision of camera calibration [2], [5].

Traditional calibration aims to determine the camera models by using calibration objects. As a result, the intrinsic parameters (focal length, axis skew, principal point, lens distortions) and the extrinsic parameters, i.e. the 6-DoF pose of each camera are determined. Intrinsic calibration is most easily done by imaging a planar pattern in different, unknown orientations [6], [7], rather than engineering a 3D calibration object. For pose estimation, near-range calibration targets are often used [8], [9]. These are not well-suited for far-range applications as the projection models are tuned for near-range and small errors in determining the arrangement may lead to significant far-range reconstruction errors [5]. Another approach is to set up a planar calibration scene up to several tens of meters [10], [5]. Although a large number of feature points are used at intrinsic calibration, the cited approaches share the drawback that all extrinsic parameters are estimated from a relatively small number of features in the far-range setup, and little is known or reported about the precision of these setups. We experienced that even with special care and laser-based measurement, small errors in laser source positions are amplified in the positions of the markers. Also, planarity assumption may easily be violated to some extent in far-range setups. In this paper, the issues just mentioned are addressed in several ways.

First, our earlier approach in [11] is revisited by giving a Maximum Likelihood (ML) formulation for the problem of stereo pose estimation. The method takes inaccuracies in the arrangement into consideration and requires at least a rough estimation of measurement uncertainties. As a side effect, the 3D structure of the marker arrangement is rectified, and discrepancies from planarity can be observed.

Second, our approach is extended by decoupling the problem into first finding the inter-camera relative pose and then determining the pose of the stereo rig with respect to the far-range arrangement. Inter-camera pose is estimated from a large number of point matches [12] incorporated into the procedure as new measurement data. The matches are established between image pairs acquired during normal operation of the vision system and the relative pose is found via computing the *essential matrix*. Such image pairs are naturally available in all applications but are rarely incorporated into the off-line calibration procedure. The essential matrix was introduced by Longuet-Higgins [13] and its computation is well studied in the computer vision literature [14], [15]. It captures the *epipolar constraint* arising from the geometry of two *calibrated* views. Its knowledge is equivalent to knowing the camera models and the 3D scene, up to a 7-DoF similarity transformation. The global scale (or, equivalently, the baseline length), the 3D orientation and the position of the stereo rig remain unknown. Methods are presented for using the far-range arrangement to reveal the remaining ambiguity. Inter-camera pose, computed from a high number of matches, is more accurate, compared to the case, when both the relative and absolute poses are retrieved from the far-range arrangement with only a few (and potentially poorly localized) features.

It will be shown that optimizing the absolute pose of the rig with fixed inter-camera pose may not fit well to the far-range data. However, if intrinsic parameters are allowed to change based on a 3D fitting criterion, and the relative pose is recomputed from the matches, good fitting can be achieved. The price that the changed intrinsic parameters do not minimize the criterion used at intrinsic calibration any more can be well compensated. The new sub-optimum is different than the one obtained without the introduced dataset, even though the latter fits well to the original "training datasets".

Theoretical background is presented in Section II. The ML formulation of pose estimation is given in Section III and the proposed decoupled approach is developed in Section IV. In Section V, justifying results on real and simulated data are presented. Finally, Section VI concludes the paper.

## II. Theoretical background

### A. Projection model per camera

Each camera is modeled with a linear pinhole projection, extended with the non-linear radial lens distortion model

$$\begin{pmatrix} u_d \\ v_d \end{pmatrix} = (1 + d_1 r^2 + d_2 r^4) \begin{pmatrix} u - c_u \\ v - c_v \end{pmatrix} + \begin{pmatrix} c_u \\ c_v \end{pmatrix}, \quad (1)$$

where $(u_d, v_d)^T$ is the distorted point, $(c_u, c_v)^T$ is the distortion center, $d_1$ and $d_2$ are the distortion coefficients, $r^2 = (u - c_u)^2 + (v - c_v)^2$, and $(u, v)^T = (X_c/Z_c, Y_c/Z_c)^T$ is the projection of the 3D point $\mathbf{M}_c = (X_c, Y_c, Z_c, 1)^T$, given in the camera reference frame. Rasterization of the distorted 2D point (or Euclidean direction) $\mathbf{d} = (u_d, v_d, 1)^T$ is modeled as

$$\mathbf{m} = \mathbf{K}\mathbf{d}, \quad (2)$$

where $\mathbf{m} = (x, y, 1)^T$ is the image of $\mathbf{M}_c$ in pixels, and the non-singular $3 \times 3$ *camera calibration matrix* $\mathbf{K}$ directly incorporates the principal point $(x_0, y_0)^T$, the axis skew $\gamma$ and the relative focal lengths $\alpha, \beta$. Thus, each camera is described by the 9 intrinsic parameters $\alpha, \beta, \gamma, x_0, y_0, d_1, d_2, c_u$ and $c_v$.

### B. Stereo projection model

Provided that radial distortion is accurately estimated, image points can be corrected. The projection models then simplify to $\mathbf{m}_l = \mathbf{P}_l\mathbf{M}$ and $\mathbf{m}_r = \mathbf{P}_r\mathbf{M}$, $l$ and $r$ standing for left and right camera. The $3 \times 4$ camera matrices $\mathbf{P}_l$ and $\mathbf{P}_r$ are

$$\begin{aligned} \mathbf{P}_l &= \mathbf{K}_l \begin{bmatrix} \mathbf{R}_l & \mathbf{t}_l \end{bmatrix} = \mathbf{K}_l \begin{bmatrix} \mathbf{R} & \mathbf{t} \end{bmatrix} \mathbf{T}_r \\ \mathbf{P}_r &= \mathbf{K}_r \begin{bmatrix} \mathbf{R}_r & \mathbf{t}_r \end{bmatrix} = \mathbf{K}_r \begin{bmatrix} \mathbf{I} & \mathbf{0} \end{bmatrix} \mathbf{T}_r \end{aligned} \quad (3)$$

where $\mathbf{R}_l, \mathbf{R}_r, \mathbf{R}$ are rotation matrices, $\mathbf{t}_l, \mathbf{t}_r, \mathbf{t}$ are translation vectors, $\mathbf{T}_r$ is a 6-DoF Euclidean transformation describing the pose of the stereo rig in the world, $\{\mathbf{R}, \mathbf{t}\}$ describes the inter-camera pose, while $\{\mathbf{R}_l, \mathbf{t}_l\}$ and $\{\mathbf{R}_r, \mathbf{t}_r\}$ are the individual absolute poses of the cameras. Note that

$$\mathbf{T}_r = \begin{bmatrix} \mathbf{R}_r & \mathbf{t}_r \\ \mathbf{0}^T & 1 \end{bmatrix}, \quad \text{and} \quad \mathbf{R} = \mathbf{R}_l\mathbf{R}_r^T, \quad \mathbf{t} = \mathbf{t}_l - \mathbf{R}\mathbf{t}_r. \quad (4)$$

Since a rotation matrix is an over-parametrization of a 3-DoF rotation, the angle-axis (Rodrigues) parametrization is used instead [15]. $\mathbf{a} \in \mathcal{R}^3$ is the direction of the rotation axis, and $||\mathbf{a}|| \in [0, \pi)$ is the rotation angle. Hence, the angle-axes $\mathbf{a}_l, \mathbf{a}_r, \mathbf{a}$ are defined for $\mathbf{R}_l, \mathbf{R}_r, \mathbf{R}$, respectively.

### C. The epipolar constraint

Corresponding points $\mathbf{m}_l \leftrightarrow \mathbf{m}_r$, observations of the same 3D point, obey the *epipolar constraint* $\mathbf{m}_l^T\mathbf{F}\mathbf{m}_r = 0$, where $\mathbf{F}$ is the rank-2, 7-DoF *fundamental matrix*. In the calibrated case, $\mathbf{K}_l$ and $\mathbf{K}_r$ are known, and gaze directions through $\mathbf{m}_l$ and $\mathbf{m}_r$ can be recovered as $\mathbf{d}_l = \mathbf{K}_l^{-1}\mathbf{m}_l$ and $\mathbf{d}_r = \mathbf{K}_r^{-1}\mathbf{m}_r$ using (2). Substitution into $\mathbf{m}_l^T\mathbf{F}\mathbf{m}_r = 0$ leads to

$$\mathbf{d}_l^T\mathbf{E}\mathbf{d}_r = 0, \qquad \mathbf{E} = \mathbf{K}_l^T\mathbf{F}\mathbf{K}_l, \quad (5)$$

where $\mathbf{E}$ is the rank-2 *essential matrix* first studied in [13]. It is not difficult to show that $\mathbf{E}$ decomposes as

$$\mathbf{E} \sim [\mathbf{t}]_\times \mathbf{R}, \quad (6)$$

where $\sim$ is equality up to scale, $[\cdot]_\times$ is the matrix operator of vector cross product, and $\{\mathbf{R}, \mathbf{t}\}$ are the relative pose parameters introduced earlier. It can be shown that $\mathbf{E}$ has 5 DoF and it has one zero and two equal singular values [15].

### D. Retrieving the cameras up to a similarity transformation

The 7-point algorithm can be used in conjunction with the RANSAC algorithm [15] to effectively eliminate false matches that do not obey the epipolar constraint. The remaining point pairs can be used to compute $\mathbf{F}$ using regression methods [14]. These are all equivalent to minimizing some cost function $\mathcal{C}_F$. We have chosen to minimize triangulation-reprojection errors with R. Hartley's optimal projective triangulation method [15].

The essential matrix can be computed from the rank-2 estimate of $\mathbf{F}$ via (5). However, this will not be a valid essential matrix with two equal singular values in practice, due to imperfection of $\mathbf{F}$. If $\mathbf{U}\,\mathrm{diag}(s_1, s_2, 0)\mathbf{V}^T$ is the SVD of the computed matrix, then $\mathbf{U}\,\mathrm{diag}(s, s, 0)\mathbf{V}^T$ with $s = (s_1 + s_2)/2$ is a valid and suitable estimate $\mathbf{E}$ of the essential matrix.

The inter-camera pose parameters $\{\mathbf{R}, \mathbf{t}\}$ can be retrieved by decomposing $\mathbf{E}$ as in (6). However, the SVD of $\mathbf{E}$ and, consequently, the decomposition (6) is not unique, because of the equal singular values. Valid solutions are

$$\mathbf{R}_{1,2} = \mathbf{U}\mathbf{W}\mathbf{V}^T \text{ or } \mathbf{U}\mathbf{W}^T\mathbf{V}^T, \qquad \mathbf{t} = \pm\lambda\mathbf{u}_3, \quad (7)$$

where $\mathbf{W}$ is a global constant matrix [15], $\mathbf{u}_3$ is the third column of $\mathbf{U}$ and $\lambda > 0$ is an unknown scale factor. Fixing $\lambda$, there are four discrete solutions of the relative arrangement. Only one of the solution meets the requirement that observed 3D points lie in front of both cameras [15]. The correct setup can be selected via triangulation of the matched point pairs using the four possible pair of camera matrices $\mathbf{K}_l \begin{bmatrix} \mathbf{R}_{1,2} & \pm\mathbf{u}_3 \end{bmatrix}$ and $\mathbf{K}_r \begin{bmatrix} \mathbf{I} & \mathbf{0} \end{bmatrix}$. The validated camera pair can be written as

$$\mathbf{P}'_l = \mathbf{K}_l \begin{bmatrix} \mathbf{R} & \mathbf{t}' \end{bmatrix}, \quad \mathbf{P}'_r = \mathbf{K}_r \begin{bmatrix} \mathbf{I} & \mathbf{0} \end{bmatrix}, \quad \mathbf{t}' = \lambda^{-1}\mathbf{t}, \quad (8)$$

where the baseline length $\lambda = ||\mathbf{t}||$ is unknown and is chosen to guarantee $||\mathbf{t}'|| = 1$. These camera matrices are related to (3) via an unknown similarity transformation $\mathbf{T}_s$ as

$$\mathbf{P}_l = \mathbf{P}'_l\mathbf{T}_s, \quad \mathbf{P}_r = \mathbf{P}'_r\mathbf{T}_s, \quad \mathbf{T}_s = \begin{bmatrix} \mathbf{R}_r & \mathbf{t}_r \\ \mathbf{0}^T & \lambda \end{bmatrix}. \quad (9)$$

Equalization of singular values of $\mathbf{E}$ comes for the price that the fundamental matrix $\mathbf{K}_l^{-T}[\mathbf{t}']_\times \mathbf{R}\mathbf{K}_r^{-1}$ does not minimize the cost $\mathcal{C}_F$ any more, and the point matches may fall far from obeying the epipolar constraint. Thus, it is necessary to re-minimize $\mathcal{C}_F$ with the inter-camera pose parameters $\{\mathbf{R}, \mathbf{t}\}$. Here, a minimal parametrization of rotation, e.g. the angle-axis vector $\mathbf{a}$ introduced earlier, should be used instead of $\mathbf{R}$.

### E. Intrinsic calibration

Intrinsic calibration is performed independently for the cameras, using a planar checkerboard pattern, with our own implementation [16] of the method of Z. Zhang [7], which is well known to be a simple method providing high accuracy. The solution is found by minimizing the reprojection error

$$\mathcal{C}_{int}(\mathbf{p}) = \sum_{j=1}^{b}\sum_{i=1}^{n_j} d^2(\mathbf{m}_i^j, \varphi(\mathbf{p}^j, \mathbf{M}_i)), \tag{10}$$

where $\mathbf{M}_i$ is the $i$-th feature point in the checkerboard, $\mathbf{m}_i^j$ is its image in the $j$-th pose of the board, $\varphi$ is the non-linear projection model, $\mathbf{p}^j$ is the vector of 9 intrinsic parameters and 6 extrinsic parameters determining the $j$-th pose of the board. Provided that $b$ board poses are used and $n_j$ extracted features are available in the $j$-th pose, there are in total $2\sum_{j=1}^{b} n_j$ measurements and $9+6b$ parameters. Note that the $6b$ extrinsic parameters (the checkerboard poses) are of no interest in the final application.

## III. Calibration without fixed inter-camera pose

### A. Initial pose estimation with respect to a plane

Once the camera calibration matrix $\mathbf{K}_k$ $(k = l, r)$ is retrieved from intrinsic calibration, it is possible to roughly determine the absolute pose $\{\mathbf{R}_k, \mathbf{t}_k\}$ of a camera with respect to a plane via computing the homography between the plane and the image, if known control points in the plane are observed with the cameras. The same method is used to determine the checkerboard poses at intrinsic calibration [7].

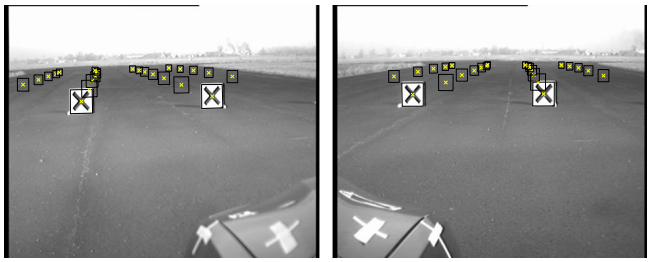

Fig. 1. Stereo pose estimation with respect to a far-range quasi-planar arrangement. 1 of 7 image pairs taken of our far-range X-marker setup. Different marker rows are on different images to avoid occlusion between markers. Computer-aided marker center extraction results are overlaid.

### B. Maximum Likelihood formulation of stereo pose estimation

It is possible to fine-tune the camera poses by taking imperfections and measurement errors of the 3D calibration arrangement into consideration. Let $n$ known control points $\tilde{\mathbf{M}}_i = (X_i, Y_i, Z_i)^T$ be given in the world frame, with detected and radially corrected images $\tilde{\mathbf{m}}_{li} = (x_{li}, y_{li})^T \leftrightarrow \tilde{\mathbf{m}}_{ri} = (x_{ri}, y_{ri})^T$. Because of independent errors in 3D localization and in feature extraction, there are no camera poses that map $\tilde{\mathbf{M}}_i$ exactly to $\tilde{\mathbf{m}}_{li}$ and $\tilde{\mathbf{m}}_{ri}$. The task is to find the corrected 3D points $\hat{\mathbf{M}}_i = (\hat{X}_i, \hat{Y}_i, \hat{Z}_i)^T$ with exact projections $\hat{\mathbf{m}}_{li} \leftrightarrow \hat{\mathbf{m}}_{ri}$ and, jointly, find the camera poses that minimize the differences between measurements and corrections. We introduce the measurement vectors $\bar{\mathbf{m}}_k = (x_{k1}, y_{k1}, \ldots, x_{kn}, y_{kn})$ with $k = l, r$, $\bar{\mathbf{M}} = (X_1, Y_1, Z_1, \ldots, X_n, Y_n, Z_n)^T$, and, similarily, the vectors $\hat{\mathbf{m}}_k$ and $\hat{\mathbf{M}}$. Suppose $\bar{\mathbf{m}}_k$ are realizations of a $2n$-dimensional stochastic vector variable with Gaussian distribution, mean $\hat{\mathbf{m}}_k$ and common variance $\sigma^2$, and $\hat{\mathbf{M}}$ is an outcome of a $3n$-dimensional variable with Gaussian distribution, mean $\hat{\mathbf{M}}$ and covariance matrix $\mathbf{\Sigma}$. We have shown in [11], that the Maximum Likelihood Estimate (MLE) of the poses and the 3D structure can be found by minimizing

$$\mathcal{C}_{ML}(\mathbf{a}_l, \mathbf{t}_l, \mathbf{a}_r, \mathbf{t}_r, \hat{\mathbf{M}}) = \tfrac{1}{\sigma^2}||\bar{\mathbf{m}} - \hat{\mathbf{m}}||_2^2 + ||\bar{\mathbf{M}} - \hat{\mathbf{M}}||_{\mathbf{\Sigma}}^2, \tag{11}$$

subject to the linear projection constraints

$$\begin{pmatrix}\hat{\mathbf{m}}_{ki}\\ 1\end{pmatrix} \sim \mathbf{K}_k \begin{bmatrix}\mathbf{R}_k & \mathbf{t}_k\end{bmatrix} \begin{pmatrix}\hat{\mathbf{M}}_i\\ 1\end{pmatrix}, \quad k = l, r. \tag{12}$$

Here, $\{\mathbf{a}_k, \mathbf{t}_k\}$, or alternatively, $\{\mathbf{R}_k, \mathbf{t}_k\}$ are the absolute poses, and $||\bar{\mathbf{M}} - \hat{\mathbf{M}}||_{\mathbf{\Sigma}}^2 = (\bar{\mathbf{M}} - \hat{\mathbf{M}})^T \mathbf{\Sigma}^{-1} (\bar{\mathbf{M}} - \hat{\mathbf{M}})$.

(11) requires exact knowledge of lens distortions, the camera calibration matrices $\mathbf{K}_l$ and $\mathbf{K}_r$, and the measurement uncertainties $\sigma^2$ and $\mathbf{\Sigma}$. In practice, only an estimation can be given to all these parameters. The inter-camera pose $\{\mathbf{R}, \mathbf{t}\}$ can be computed from $\mathbf{R}_k$ via (4).

## IV. Calibration with fixed inter-camera pose

In this section, it is supposed that the inter-camera pose $\{\mathbf{R}, \mathbf{t}\}$ is already computed and optimized via the essential matrix from a large number of point correspondences established between views, as described in Section II-D. The task is to find the similarity transformation $\mathbf{T}_s$ introduced in Equation (9), i.e. to find the absolute pose $\{\mathbf{R}_r, \mathbf{t}_r\}$ of the rig and the baseline length $\lambda = ||\mathbf{t}||$. Knowledge of $\lambda$ is equivalent to knowing the overall scale of a potential reconstruction. As in this scenario, the point matches are taken from image pairs acquired in normal operation of the stereo system (e.g. in traffic), and the far-range planar calibration scene has not been exploited yet at all, the latter will be used to determine $\mathbf{T}_s$.

### A. Two-stage initial pose estimation by reprojection

Since the absolute pose of the rig is defined as the absolute pose of the right camera, the parameters $\{\mathbf{R}_r, \mathbf{t}_r\}$ can be determined simply with the pose estimation method in Section III-A. In this procedure, only the features $\mathbf{m}_{ri}$ detected in the right image and the 3D features $\mathbf{M}_i$ localized on the road plane are used. The resulted pose can be further refined by minimizing the reprojection error

$$\mathcal{C}_r(\mathbf{a}_r, \mathbf{t}_r) = \sum_{i=1}^{n} d^2(\mathbf{m}_{ri}, \mathbf{P}_r \mathbf{M}_i), \tag{13}$$

where $d(\cdot,\cdot)$ is Euclidean distance in pixels, $\mathbf{P}_r$ is defined in (3) and $\mathbf{a}_r$ is related to $\mathbf{R}_r$ via the Rodrigues-formula.

The remaining task is to determine the baseline length $\lambda$ by taking into consideration the features $\mathbf{m}_{li}$, detected projections of $\mathbf{M}_i$ in the left image. Here, a modification of the pose-from-plane technique in Section III-A is used.

Suppose $\mathbf{M}_i$ is given in the XZ-plane of the world reference frame, that is $\mathbf{M}_i = (X_i, 0, Z_i, 1)^T$, and introduce $\mathbf{m}_{xzi} = (X_i, Z_i, 1)^T$. From (8) and (9), the camera matrix $\mathbf{P}_l$ is

$$\mathbf{P}_l = \mathbf{K}_l\mathbf{R}\begin{bmatrix}\mathbf{R}_r & \mathbf{t}_r + \lambda\mathbf{R}^T\mathbf{t}'\end{bmatrix}, \tag{14}$$

where only $\lambda$ is unknown at this stage. The homography between the plane and its image can be determined from $\mathbf{m}_{li} \sim \mathbf{P}_l\mathbf{M}_i \sim \mathbf{H}\mathbf{m}_{xzi}$. It follows that

$$\begin{bmatrix}\mathbf{h}_1 & \mathbf{h}_2 & \mathbf{h}_3\end{bmatrix} = \eta\mathbf{K}_l\mathbf{R}\begin{bmatrix}\mathbf{r}_1 & \mathbf{r}_3 & \mathbf{t}_r + \lambda\mathbf{R}^T\mathbf{t}'\end{bmatrix}, \tag{15}$$

where $\eta$ is an unknown factor, $\mathbf{h}_j$ and $\mathbf{r}_j$ are the $j$-th column of $\mathbf{H}$ and $\mathbf{R}_r$, respectively. $\mathbf{H}$ can be computed from at least four correspondences $\mathbf{m}_{li} \leftrightarrow \mathbf{m}_{xzi}$ [15].

$\eta$ is over-determined by $\mathbf{A} = \eta\mathbf{B}$, where $\mathbf{A} = [a_{pq}] = \mathbf{R}^T\mathbf{K}_l^{-1}\begin{bmatrix}\mathbf{h}_1 & \mathbf{h}_2\end{bmatrix}$ and $\mathbf{B} = [b_{pq}] = \begin{bmatrix}\mathbf{r}_1 & \mathbf{r}_3\end{bmatrix}$. However, equation (15), and, consequently, $\mathbf{A} = \eta\mathbf{B}$ is not satisfied exactly in practice, since known parameters are error-prone. It is reasonable to minimize the Frobenius-norm $||\mathbf{A} - \eta\mathbf{B}||_F$.

$$\hat{\eta} = \text{argmin}_\eta \sum_{p,q}(a_{pq} - \eta b_{pq})^2 = \frac{\Sigma_{ij}(\mathbf{A} \circ \mathbf{B})}{||\mathbf{B}||_F^2}, \tag{16}$$

where $\Sigma_{ij}(\cdot)$ is summation over all indices and $\circ$ is the element-wise (Hadamard) product.

Having determined $\hat{\eta}$, $\lambda$ can be calculated from (15), or, $\mathbf{R}^T\mathbf{K}_l^{-1}\mathbf{h}_3 = \eta(\mathbf{t}_r + \lambda\mathbf{R}^T\mathbf{t}')$. Again, this is not satisfied exactly and it is resonable to re-apply (16), which results

$$\hat{\lambda} = \mathbf{t}'^T(\eta^{-1}\mathbf{K}_l^{-1}\mathbf{h}_3 - \mathbf{R}\mathbf{t}_r). \tag{17}$$

The scale $\hat{\lambda}$ can be fine-tuned by minimizing the reprojection error $\mathcal{C}_l(\lambda)$ in the left image, similarily to (13).

### *B. Pose estimation by triangulation and 3D registration*

The method just presented is best to use with neglectable errors in the world points, as it reduces all errors to the images. The reverse approach is to model all errors in 3D. Starting from the image correspondences $\mathbf{m}_{li} \leftrightarrow \mathbf{m}_{ri}$ arising from the 3D locations $\mathbf{M}_i$ of the markers in the far-range arrangement, the points $\mathbf{M}_i$ can be reconstructed up to a similarity transformation via triangulation with camera matrices $\mathbf{P}'_l \leftrightarrow \mathbf{P}'_r$ in (8). This is because $\mathbf{m}_{li} = \mathbf{P}_l\mathbf{M}_i = (\mathbf{P}_l\mathbf{T}_s^{-1})(\mathbf{T}_s\mathbf{M}_i) = \mathbf{P}'_l\hat{\mathbf{M}}_i$, and similarily to $\mathbf{m}_{ri}$. Thus, ideally, the reconstructed points $\hat{\mathbf{M}}_i$ are related to the original points $\mathbf{M}_i$ via

$$\hat{\mathbf{M}}_i = \mathbf{T}_s\mathbf{M}_i, \tag{18}$$

and the estimation of the pose $\{\mathbf{R}_r, \mathbf{t}_r\}$ and scale $\lambda$ in $\mathbf{T}_s$ leads to a 3D point cloud registration problem. A straightforward solution is to minimize

$$\mathcal{C}_{3D}(\mathbf{a}_r, \mathbf{t}_r, \lambda) = \sum_{i=1}^{n} d^2(\mathbf{M}_i, \hat{\mathbf{M}}_i), \tag{19}$$

where $\mathbf{a}_r$ is the angle-axis vector corresponding to $\mathbf{R}_r$.

### *C. Maximum likelihood with fixed inter-camera pose*

It is possible to derive the MLE of the absolute pose of the stereo rig with fixed inter-camera pose, provided that the inter-camera pose $\{\mathbf{R}, \mathbf{t}\}$ is known exactly and provided that the suppositions of Section III-B hold. In this case, the cost function $\mathcal{C}_{ML}(\mathbf{a}_l, \mathbf{t}_l, \mathbf{a}_r, \mathbf{t}_r, \hat{\mathbf{M}})$ in (11) is replaced by $\mathcal{C}^{rel}_{ML}(\mathbf{a}, \mathbf{t}, \mathbf{a}_r, \mathbf{t}_r, \hat{\mathbf{M}})$, and the pose of the left camera $\{\mathbf{a}_l, \mathbf{t}_l\}$ is implicitely computed during minimization, from the rig pose $\{\mathbf{a}_r, \mathbf{t}_r\}$ and the inter-camera pose $\{\mathbf{a}, \mathbf{t}\}$.

## V. Results

### *A. Real stereo setup*

Real experiments were performed using two analog cameras fixed to the side mirrors of a car. The cameras had nominally 8 mm focal length, $33.4°/25.4°$ horizontal/vertical FoV, $480 \times 384$ pixels resolution, $10 \times 9.375$ $\mu$m pixel sizes and 1.067 pixel aspect ratio. A world reference frame has been designated on the ground, right under the front license plate, $X$ axis pointing to the left, $Z$ forward and $Y$ up, XZ roughly being the road plane. In this frame, nominal location of the cameras are $(\pm 1.00, 1.13, -1.52)^T$ meters ($\pm 10$ cm).

### *B. Intrinsic calibration*

Intrinsic checkerboard-based calibration have been performed separately for the two cameras as described in Section II-E, with ccal GUI, our Matlab Camera Calibration Toolbox [16]. In total, 1190/1232 features were extracted with a corner detector in 16/16 poses of the checkerboard. The residual reprojection errors are characterized by $\sigma = 0.287$ and 0.226 pixels, respectively, for the left/right camera. It has been found that fixing the skew $\gamma = 0$ and aspect ratio $\beta/\alpha = 1.067$ increases $\sigma$ by only 5% to 0.300 pixels for the left, and by 13% to 0.256 pixels for the right camera but nulling out radial distortion implies an increase of 57% and 86%, respectively. Radial displacement predicted by the model reach 14 pixels at the corners. Thus, radial distortion can not be neglected. In turn, we keep the constraints $\gamma = 0$ and $\beta/\alpha = 1.067$, as these are important for autocalibration, and we are planning to evaluate on-line calibration techniques in the stereo context [15], [17]. Resulted focal lengths are $\alpha = 785.1 \pm 4.5$ and $784.0 \pm 3.8$ pixels, principal points are $(205.8 \pm 12.0, 194.2 \pm 9.6)^T$ and $(239.2 \pm 9.7, 170.0 \pm 7.9)^T$. The $\pm$ values are 99% confidence intervals estimated by back-propagation of covariance from the image to the parameters.

### *C. Pose estimation without fixed inter-camera pose*

A quasi-planar arrangement of 24 marker plates of size $50 \times 50$ cm, each with an X-pattern, was installed, up to 40 meters in front of the car, as shown in Figure 1. Markers were localized with laser distance meters from two reference points on the two sides of the car, with a baseline of 15 m. Not only $(X, Y, Z)$ marker center locations were computed, but their uncertainty is characterized, as well, by Monte-Carlo simulation of the marker localization procedure (see [11] for more details). The $72 \times 72$ covariance matrix $\mathbf{\Sigma}$ of all 3D coordinates is computed from the 24 3D point clusters. The

X-sizes of 99% covariance ellipsoids per marker range from 60 to 150 cm and increase with distance, while the $Z$ sizes range from 6 to 45 cm, *decreasing with distance*.

In the images, the centers of the X-markers were localized within an accuracy of 0.5 pixels, by selecting and adjusting the four corners of the plates manually, a procedure helped by a Matlab-program that continously fitted a perspectively distorted model of the X-shape to the selected corners. Note that measurements in the images are much more precise than 3D measurements: the 3D Monte-Carlo clouds project into 2D clouds comparable to the size of the imaged markers.

Next, the procedure in Section III is performed and camera poses are found by minimizing the ML cost (11). Encouraging results are received (Fig.2 and "ML" in Table I). Residuals in 3D are less then 18 cm, and in the images, they are within $\pm 0.4$ pixels per coordinate. Optimal camera locations are $(0.95, 1.12, -1.57)^T$ m and $(-1.02, 1.10, -1.53)^T$ m (within $\pm 5$ cm from nominal locations measured with tape, see Section V-A). The optimal baseline length is $\hat{\lambda}_{ML} = 1.973$ m.

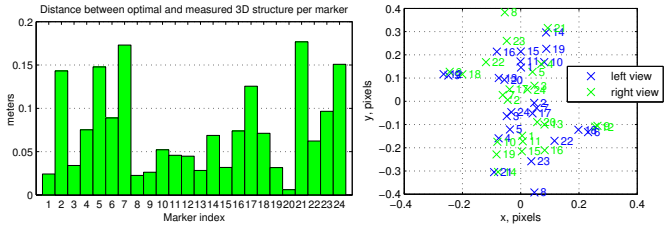


Fig. 2. Residual errors of ML pose estimation per marker after minimizing (11). Left: 3D distances between the measured ($\mathbf{M}_i$) and the optimal marker locations ($\hat{\mathbf{M}}_i$). Right: 2D residuals in the images. See "ML" in Table I

### D. Rig pose and scale estimation with fixed inter-camera pose

26 image pairs were selected from video streams acquired in traffic. In total, 1943 pairs of SIFT features were matched, with the Matlab-based SIFT implementation of A. Vedaldi [18]. False matches violating the epipolar constraint were removed automatically by RANSAC (see Fig.3) and all remaining false matches were removed manually. Valid matches were used to compute the inter-camera relative pose via the essential matrix, as described in Section II-D. Here, optimization is indispensable, since singular value equalization of $\mathbf{E}$ severely deteriorates the fitting of the matches to the epipolar constraint. With optimization, the RMS epipolar residual is 0.42 pixels.

First, the ML method "trained on the X-marker dataset" was evaluated using the new dataset. Surprisingly, epipolar errors arising from the MLE are high and biased (Fig.5). This may be an indication of lower accuracy in far-range reconstruction, even though the fitting to the "training data" was good (Fig.2).

Next, the methods of Section IV-A and IV-B were used to find the absolute pose of the rig, with fixed inter-camera pose computed directly from the matches. Both methods resulted in a bad fit to the X-marker data, although optimizations converged securely. See residuals under "reproj." and "3D reg." in Table I. In the first case, reprojection errors were up to 6 pixels in the right image after determinig the pose $\{\mathbf{R}_r, \mathbf{t}_r\}$, and 14 in the left image after determining scale $\lambda$.

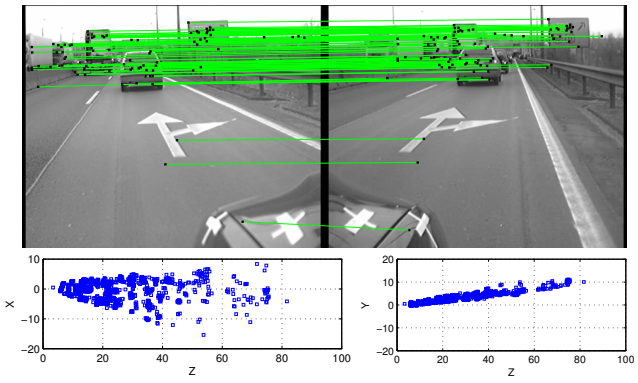


Fig. 3. The new dataset used for calibration: 1 of 26 image pairs from videos in traffic (top). SIFT matches are established and those violating the epipolar constraint are removed by RANSAC. Remaining false matches (e.g. those at the bottom) are eliminated manually. A reconstruction of all 1943 SIFT-matches in the right camera frame after relative pose computation (bottom). The reconstruction has nominal scale (units are in meters).

An additional joint reprojection error minimization with both $\{\mathbf{R}_r, \mathbf{t}_r\}$ and $\lambda$ only changed the distribution of errors in the images (8-8 pixels).

In the second case, the 3D point clouds were registered with individual residual 3D errors up to 1.4 meters. In Figure 4, it can be seen that systematic errors are resulted that exceed the marker sizes. This clearly can not be caused by independent feature extraction errors of within 0.5 pixels.

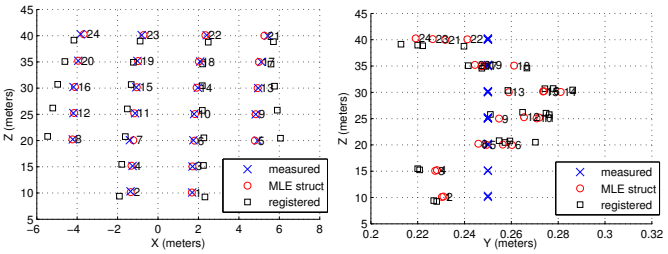


Fig. 4. The laser-measured 3D calibration scene with planarity hypothesis, the Maximum Likelihood optimum structure and the structure via 3D registration (fixed inter-camera pose). A hump in the YZ-profile of the structure was observed in reality, as well. Note systematic errors in the 3D-registered structure on the left. See numerical results in "3D reg." column of Table I.

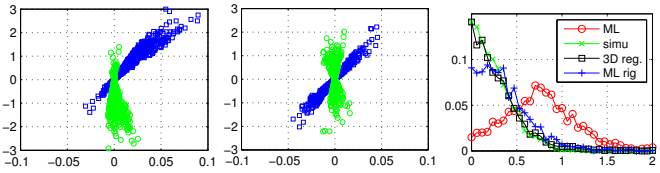


Fig. 5. Point–epipolar line residuals for 1943 SIFT matches: ML method not based on the matches (left), methods computing the inter-camera pose from the matches (center). These errors are only affected by camera intrinsics and the inter-camera pose. Histogram of the errors for different methods (right).

### E. Investigation of inconsistencies with simulations

Several Matlab-simulations have been run to determine what causes systematic errors in 3D registration and why the camera models with fixed inter-camera pose, computed from a high number of point matches, do not fit to the X-marker dataset. The optimal structure and cameras received from the ML minimizer of (11) were taken as ground truth, and all 1943

SIFT matches were corrected to perfectly obey the epipolar constraint. As a result, a perfectly consistent virtual dataset is received. Then the whole pose estimation chain, with essential matrix computation, triangulation and 3D registration (see Section IV-B) is run. With no errors in the data, the final fitting, including 3D registration is perfect. First, Gaussian random noise with $\sigma = 0.5$ pixels is introduced only in SIFT feature locations. This is very pessimistic, as epipolar errors having similar distribution to our real errors can be reproduced with $\sigma = 0.3$ ("simu" in Fig.5), while the distribution for $0.3 \pm 0.05$ differs significantly. With the pessimistic amount of noise, the mean of all RMS 3D registration residuals from 100 trials is 0.14 m, while their maximum is 0.42 m. The latter is suprisingly high, considering the high number of matches, but is still only a fraction of 0.95 m experienced with real data (see Table I). Thus, errors in feature locations are unlikely to be primary reason for our high 3D registration errors.

In turn, when one of the intrinsics $\alpha, x_0, y_0$ is perturbed, the high errors in Fig.4 could be reproduced. Roughly using the 99% confidence intervals (Section V-B), $\alpha$ was perturbed with $\pm 4$ pixels, while $x_0, y_0$ were perturbed with $\pm 8$ pixels. 3D registration errors ranged between 0.30 and 1.44 meters in different scenarios. This justifies that small imprecision in the intrinsics may have caused the large systematic errors in far-range reconstruction. Inversely, the far-range setup may be used to constrain intrinsic parameters.

TABLE I
RMS RESIDUAL ERROR NORMS IN THE DIFFERENT DATASETS FOR DIFFERENT POSE ESTIMATION METHODS. HIGHEST ERRORS ARE IN BOLD.

| | | ⟵inter-camera pose from matches⟶ | | | |
|---|---|---|---|---|---|
| | ML | reproj. | 3D reg. | intr-to-X | ML rig |
| checker,left (pix) | 0.300 | 0.300 | 0.300 | 0.309* | 0.309* |
| checker,right (pix) | 0.256 | 0.256 | 0.256 | 0.264** | 0.264** |
| X 3D (meters) | 0.091 | 0 | **0.952** | 0.144 | 0.089 |
| X images (pix) | 0.21 | **3.60** | 0.44 | 0.38 | 0.39 |
| SIFT,epipolar (pix) | **0.90** | 0.42 | 0.42 | 0.49 | 0.49 |

*__5.7__ and **__17.3__ before optimized repositioning of the checkerboards

### *F. Achieving consistency over all real datasets*

Inspired by the previous idea, $\alpha, x_0, y_0$ were optimized for both cameras to minimize 3D registration errors in the far-range X-arrangement, using real data. Convergence is achieved with a change of $-5.8$ and $-5.2$ pixels in $\alpha$, and $(+0.6, +1.6)$ and $(-1.9, +2.8)$ in $(x_0, y_0)$. Changes in intrinsics come for the price, that reprojection errors of the checkerboard features, used preliminarily at intrinsic calibration, become biased (see remark under Table I, column "intr-to-X"). Fortunately, checkerboard residuals can be resumed to near their original amount (with an increase of only +3%, see Table I) by re-minimizing (10) with only the $6 \times 16$ extrinsic parameters independently for the two cameras (new intrinsics fixed). Hence, intrinsic parameter changes are almost perfectly compensated by checkerboard repositionings, while checkerboard poses are of no interest in the final application.

Results for the ML method described in Section IV-C, with the new intrinsics and inter-camera pose fixed, are also shown in Table I ("ML rig"). Compared to "ML", it reshapes residuals by favouring the 1943 matches vs. the 24 marker projections.

## VI. CONCLUSION

Several methods for calibrating a stereo rig for far-range have been presented. Relying on a small number of noisy markers with a full parameter set to be estimated should be avoided by incorporating stable stereo matches from on-line images into the off-line pose estimation procedure. It has been shown that the maximum likelihood poses estimated from only the far-range arrangement yield significant bias in left-right consistency of the established matches. However, if inter-camera pose is determined reliably from these, only the rig-to-world pose remains subject to a small number of markers in the far-range arrangement. Consistency between the inter-camera pose and the far-range arrangement has been achieved by using the latter to constrain the intrinsic parameters, which were identified to be primary source of errors, even after a sound intrinsic calibration. Future work consists of developing a cost function that unifies all available measurements, and evaluating stereo autocalibration methods based on our results.


## ACKNOWLEDGEMENT

This work has been supported by the Hungarian Scientific Research Fund (OTKA), grant number TS-73496.


## REFERENCES


[1] M. Bertozzi et al., "Artificial vision in road vehicles," *Proceedings of the IEEE*, vol. 90, no. 7, pp. 1258–1271, 2002.

[2] Z. Sun, G. Bebis, and R. Miller, "On-road vehicle detection using optical sensors: A review," in *Proc. on the 7th International IEEE Conf. on Intelligent Transportation Systems*, 2004, pp. 585–590.

[3] S. Nedevschi et al, "Driving environment perception using stereovision," in *Proc. on IEEE Intelligent Vehicles Symposium*, 2005, pp. 331–336.

[4] M. M. Trivedi, T. Gandhi, and J. McCall, "Looking-in and looking-out of a vehicle: Computer-vision-based enhanced vehicle safety," *IEEE Trans. on Intell. Transportation Systems*, vol. 8, no. 1, pp. 108–120, 2007.

[5] T. Marita et al., "Camera calibration method for far range stereovision sensors used in vehicles," in *Intell. Vehicles Symp.*, 2006, pp. 356–363.

[6] P. F. Sturm and S. J. Maybank, "On plane-based camera calibration: A general algorithm, singularities, applications," in *IEEE Conf. on Computer Vision and Pattern Recognition*, vol. 1, 1999, pp. 432–437.

[7] Z. Zhang, "A flexible new technique for camera calibration," *IEEE PAMI*, vol. 22, no. 11, pp. 1330–1334, nov 2000.

[8] N. Kaempchen, U. Franke, and R. Ott, "Stereo vision based pose estimation of parking lots using 3d vehicle models," in *Intelligent Vehicle Symposium*, vol. 2, 2002, pp. 459–464.

[9] M. Bellino, Y. de Meneses, S. Kolski, and J. Jacot, "Calibration of an embedded camera for driver-assistant systems," in *IEEE Proc. Intelligent Transportation Systems*, Sep. 2005, pp. 354–359.

[10] A. Broggi, M. Bertozzi, and A. Fascioli, "Self-calibration of a stereo vision system for automotive applications," in *IEEE Intl. Conf. on Robotics and Automation*, vol. 4, 2001, pp. 3698–3703.

[11] A. Bódis-Szomorú, T. Dabóczi, and Z. Fazekas, "Calibration and sensitivity analysis of a stereo vision-based driver assistance system," in *Stereo Vision*, A. Bhatti, Ed. InTech, Nov. 2008, pp. 1–26.

[12] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *International Journal of Computer Vision*, vol. 60, pp. 91–110, 2004.

[13] H. C. Longuet-Higgins, "A computer algorithm for reconstructing a scene from two projections," *Nature*, vol. 293, pp. 133–135, 1981.

[14] Z. Zhang, "Determining the epipolar geometry and its uncertainty - a review," *Intl. Journal of Comp. Vision*, vol. 27, no. 2, pp. 161–195, 1998.

[15] R. Hartley and A. Zissermann, *Multiple View Geometry in Computer Vision, Second Edition*. Cambridge University Press, 2006.

[16] ccal Camera Calibration GUI Toolbox for Matlab by A. Bódis-Szomorú http://www.mit.bme.hu/∼bodis/ccalgui.html.

[17] R. Horaud, G. Csurka, and D. Demirdijian, "Stereo calibration from rigid motions," *IEEE PAMI*, vol. 22, no. 12, pp. 1446–1452, 2000.

[18] A. Vedaldi, "An open implementation of the SIFT detector and descriptor," UCLA CSD, Tech. Rep. 070012, 2007.