

MLPNP - A REAL-TIME MAXIMUM LIKELIHOOD SOLUTION TO THE PERSPECTIVE-N-POINT PROBLEM

S. Urban, J.Leitloff, S.Hinz

Institute of Photogrammetry and Remote Sensing, Karlsruhe Institute of Technology Karlsruhe
Englerstr. 7, 76131 Karlsruhe, Germany - (steffen.urban, jens.leitloff, stefan.hinz)@kit.edu
<http://www.ipf.kit.edu>

KEY WORDS: pose estimation, perspective-n-point, computer vision, photogrammetry, maximum-likelihood estimation

ABSTRACT:

In this paper, a statistically optimal solution to the Perspective-n-Point (PnP) problem is presented. Many solutions to the PnP problem are geometrically optimal, but do not consider the uncertainties of the observations. In addition, it would be desirable to have an internal estimation of the accuracy of the estimated rotation and translation parameters of the camera pose. Thus, we propose a novel maximum likelihood solution to the PnP problem, that incorporates image observation uncertainties and remains real-time capable at the same time. Further, the presented method is general, as it works with 3D direction vectors instead of 2D image points and is thus able to cope with arbitrary central camera models. This is achieved by projecting (and thus reducing) the covariance matrices of the observations to the corresponding vector tangent space.

1. INTRODUCTION

The goal of the PnP problem is to determine the absolute pose (rotation and translation) of a calibrated camera in a world reference frame, given known 3D points and corresponding 2D image observations. The research on PnP has a long history in both the computer vision and the photogrammetry community (here called camera or space resectioning). Hence, we would first like to emphasize the differences between the definitions used in both communities.

PnP Classically in literature, basically two definitions of the problem exist (Hu and Wu, 2002). In the first, the *distance based* definition, the problem is formulated in terms of the distances from the projective center to each 3D points, e.g. leading to minimal P3P solutions (Haralick et al., 1991). The second definition is *transformation based*. Here, the task is to determine the 3D rigid body transformation between object-centered and camera centered coordinate systems, e.g. (Fiore, 2001, Horaud et al., 1989). In both PnP definitions, however, the camera is always assumed to be calibrated and known, allowing to transform image measurements into unit vectors, that point from the camera to the 3D scene points. More recent approaches extend this classical definition by including unknown camera parameters into the formulation such as the focal length (Wu, 2015) or radial distortion (Kukelova et al., 2013).

Camera resectioning The task in the photogrammetric definition of the problem is to find the projection matrix $\lambda \mathbf{u} = \mathbf{P}_{3 \times 4} \mathbf{X}$ that transforms homogeneous 3D scene points \mathbf{X} to homogeneous 2D image points \mathbf{u} (Hartley and Zisserman, 2003, Luhmann et al., 2006). The projection matrix is given by $\mathbf{P}_{3 \times 4} = \mathbf{K}[\mathbf{R}|\mathbf{t}]$ and hence contains the camera matrix \mathbf{K} as well as the rotation \mathbf{R} and translation \mathbf{t} .

Thus, fundamentally, the main difference between the definitions in both communities is, that PnP solves only for the absolute pose of the same, calibrated camera. In camera resectioning the camera is assumed to be unknown and thus part of the formulation of the problem. In this paper, we assume the camera to be calibrated and known, thus we present a solution to the Perspective-n-Point problem.

Still, the need for efficient and accurate solutions is driven by a large number of applications. They range from localization of robots and object manipulation (Choi and Christensen, 2012, Collet et al., 2009), to augmented reality (Müller et al., 2013) implementations running on mobile devices and having only limited resources, thus focusing on fast solutions. Especially in industry, surveying or medical environments, involving machine vision, (close-range) photogrammetry (Luhmann et al., 2006), point-cloud registration (Weinmann et al., 2011) and surgical navigation (Yang et al., 2015), methods are demanded, that are not only robust, but also return a measure of reliability.

Even though the research on the PnP problem has a long history, few work has been published on efficient real-time solutions, that take the observation uncertainty into account. Most algorithms focus on geometric but not statistic optimality. To the best of our knowledge, the only work, that includes observation uncertainty into their framework is the Covariant EPPnP of Ferraz et al. (Ferraz et al., 2014a).

In this paper, we propose a novel formulation of a Maximum Likelihood (ML) solution to the PnP problem. Further, a general method for variance propagation from image observations to bearing vectors is exploited to avoid singular covariance matrices. In addition, we benchmark our real-time method against the state-of-the-art and show on a ground truth tracking data set, how our statistical framework can be used, to get a measure of the accuracy of the unknowns given the knowledge about the uncertain observations.

2. RELATED WORK

The minimal number of points to solve the PnP problem is three. Closed-form solutions to that minimal configuration return up to four solutions, and a fourth point can be used for disambiguation. Prominent solutions to the P3P problem requiring exactly three points are (Kneip et al., 2011),(Li and Xu, 2011) and (DeMenthon and Davis, 1992). The stability of such algorithms under noisy measurements is limited, hence they are predominately employed in RANSAC (Fischler and Bolles, 1981) based outlier rejection schemes. Apart from the P3P methods, the P4P (Fischler and

Bolles, 1981, Triggs, 1999) and P5P (Triggs, 1999) algorithms exist, still dependent on a fixed number of points.

Most solvers, however, can cope with an arbitrary number of feature correspondences. Basically, they can be categorized into iterative, non-iterative or polynomial, non-polynomial solvers. Table 1 lists all methods that are in addition evaluated in the experimental section.

Iterative solutions use different objective functions, that are minimized. In case of LHM (Lu et al., 2000) the pose is initialized using a weak perspective assumption. Then they proceed by iteratively minimizing the object space error, i.e. orthogonal deviations between observed ray directions and the corresponding object points in the camera frame. In the Procrustes PnP (Garro et al., 2012), the error between the object and the back-projected image points is minimized. The back-projection is based on the iteratively estimated transformation parameters. As iterative methods usually are only guaranteed to find local minima, (Schweighofer and Pinz, 2008) reformulated the PnP problem into a semidefinite programme (SDP). Despite its $O(n)$ complexity, the runtime of the global optimization method remains tremendous.

Likewise, early non-iterative solvers were computational demanding, especially for large point sets. Among them (Ansar and Daniilidis, 2003) with $O(n^8)$, (Quan and Lan, 1999) with $O(n^5)$ and (Fiore, 2001) with $O(n^2)$. The first efficient non-iterative $O(n)$ solution is the EPnP by (Moreno-Noguer et al., 2007), that was subsequently extended by (Lepetit et al., 2009), employing a fast iterative method to improve the accuracy. The efficiency comes from the reduction of the PnP problem to finding the position of four control points that are a weighted sum of all 3D points. After obtaining a linear solution, the weights of the four control points is refined using Gauss-Newton optimization.

The most recent, non-iterative state-of-the-art solutions are all polynomial solvers: The Robust PnP (RPnP) (Li et al., 2012) first splits the PnP problem into multiple P3P problems, that result in a set of fourth order polynomials. Then the squared sum of those polynomials as well as its derivative is calculated and the four stationary points are obtained. The final solution is selected as the stationary point with the smallest reprojection error. In the Direct-Least-Squares (DLS) (Hesch and Roumeliotis, 2011) method, a nonlinear object space cost function is formulated and polynomial resultant techniques are used to recover the (up to 27) stationary points of a polynomial equation system of fourth order polynomials. A drawback of this method is the parametrization of the rotation in the cost function by means of the Cayley parameters. To overcome this, the Accurate and Scalable PnP (ASnPnP) (Zheng et al., 2013b) and the Optimal PnP (OPnP) (Zheng et al., 2013a) use quaternion based representation of the rotation matrix. Subsequently, the (up to 40) solutions are found using the Gröbner basis technique on the optimality conditions of the algebraic cost function.

The linear, non-iterative Unified PnP (UPnP) (Kneip et al., 2014) goes one step further and integrates the solution to the NPnP (Non-Perspective-N-Point) problem. The DLS formulation is extended to include non-central camera rays and the stationary points of the first order optimality conditions on the sum of object space errors are found using the Gröbner basis methodology.

Thus far, all methods assume, that the observations are equally accurate and free of erroneous correspondences. The first PnP method, that includes an algebraic outlier rejection scheme within the pose estimation, is an extension of the EPnP algorithm called Robust Efficient Procrustes PnP (REPPnP) (Ferraz et al., 2014b).

Outliers are removed from the data by sequentially eliminating correspondences that exceed a threshold on an algebraic error. The procedure remains efficient, as the algorithm operates on the linear system, spanned by the virtual control points of EPnP and thus avoids recalculating the full projection equation in each iteration. After removing the outliers, the final solution is attained by iteratively solving the closed-form Orthogonal Procrustes problem.

Yet another extension to the aforementioned EPnP is termed Covariant EPnP (CEPPnP) (Ferraz et al., 2014a). It is the first algorithm to inherently incorporate observation uncertainty into the framework. Again the linear control point system of EPnP is formulated. Then the covariance information of the feature points is transformed to that space using its Jacobian. Finally, the Maximum Likelihood (ML) minimization is approximated by an unconstrained Sampson error.

In this paper, we propose a new real-time capable, statistically optimal solution to the PnP problem leveraging observation uncertainties. We propagate 2D image uncertainties to general bearing vectors and show, how a linear ML solution with non-singular covariance matrices can be obtained by using the reduced observation space presented by Förstner (Förstner, 2010). In contrast to the CEEPnP, where the ML estimator is used, to obtain an estimation of the control point subspace, we optimize directly over the unknown quantities, i.e. rotation and translation and thus, directly obtain pose uncertainties. Finally, we compare the results of our algorithm to a ground truth trajectory and show, that the estimated pose uncertainties are very close to the ground-truth.

| | iter. | polyn. | using Σ | reference |
|----------|-------|--------|----------------|-------------------------------|
| LHM | X | | | (Lu et al., 2000) |
| EPnP+GN | (X) | | | (Lepetit et al., 2009) |
| DLS | | X | | (Hesch and Roumeliotis, 2011) |
| RPNP | | X | | (Li et al., 2012) |
| PPnP | X | | | (Garro et al., 2012) |
| OPnP | | X | | (Zheng et al., 2013a) |
| ASnPnP | | X | | (Zheng et al., 2013b) |
| (R)EPPnP | X | | | (Ferraz et al., 2014b) |
| CEPPnP | (X) | | X | (Ferraz et al., 2014a) |
| UPnP | | X | | (Kneip et al., 2014) |
| MLPnP | (X) | | X | this |

Table 1: Comparison of all tested methods. The methods are categorized into being iterative or a polynomial solver and if they incorporate measurement uncertainty Σ . (X) depicts methods, that refine a preliminary result iteratively.

3. MLPNP

The task of Perspective-n-Point (PnP) is to find the orientation $\mathbf{R} \in SO(3)$ and translation $\mathbf{t} \in \mathbb{R}^3$ that maps the I world points $\mathbf{p}_i, i = 1, \dots, I$ to their corresponding observations \mathbf{v}_i in the camera frame. This relation is given by:

$$\lambda_i \mathbf{v}_i = \mathbf{R} \mathbf{p}_i + \mathbf{t} \quad (1)$$

where λ_i are the depths of each point and the observations \mathbf{v}_i are the measured and thus uncertain quantity having unit length, i.e. $\|\mathbf{v}_i\| = 1$. In the following the methodology of our algorithm is explained. The parameters and observations are depicted in Fig. 1.

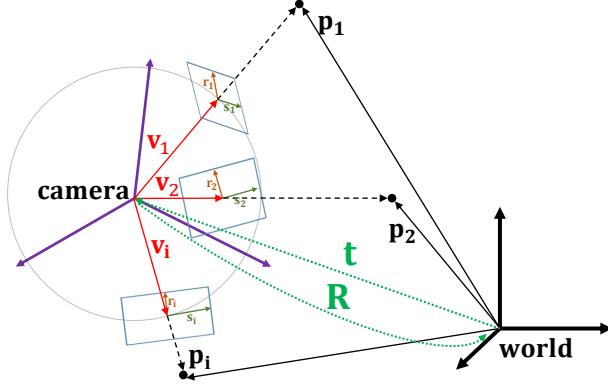


Figure 1: Observations of object points from a camera. The planes at each bearing vector \mathbf{v}_i are spanned by its respective null space vectors \mathbf{r}_i and \mathbf{s}_i .

3.1 Observations and uncertainty propagation

Let a world point $\mathbf{p}_i \in \mathbb{R}^3$ be observed by a calibrated camera and let the uncertain observation in the image plane be:

$$\mathbf{x}' = \begin{bmatrix} x' \\ y' \end{bmatrix}, \Sigma_{\mathbf{x}'\mathbf{x}'} = \begin{bmatrix} \sigma_{x'}^2 & \sigma_{x'y'} \\ \sigma_{y'x'} & \sigma_{y'}^2 \end{bmatrix} \quad (2)$$

where the uncertainty about the observed point is described by the 2D covariance matrix $\Sigma_{\mathbf{x}'\mathbf{x}'}$. Assuming an arbitrary interior orientation parametrization, the image point \mathbf{x}' is projected to its corresponding three dimensional direction in the camera frame, using the forward projection function π (e.g. \mathbf{K}^{-1} in the perspective case):

$$\mathbf{x} = \pi\mathbf{x}' = \begin{bmatrix} x \\ y \\ 1 \end{bmatrix}, \mathbf{J}_\pi = \begin{bmatrix} \frac{\partial \pi_{x'}}{\partial x'} & \frac{\partial \pi_{x'}}{\partial y'} \\ \frac{\partial \pi_{y'}}{\partial x'} & \frac{\partial \pi_{y'}}{\partial y'} \\ 0 & 0 \end{bmatrix} \quad (3)$$

with \mathbf{J}_π being the Jacobian of the forward projection. Thus, the uncertainty of the image point \mathbf{x}' is propagated using:

$$\Sigma_{\mathbf{x}\mathbf{x}} = \mathbf{J}_\pi \Sigma_{\mathbf{x}'\mathbf{x}'} \mathbf{J}_\pi^T = \begin{bmatrix} \sigma_x^2 & \sigma_{xy} & 0 \\ \sigma_{yx} & \sigma_y^2 & 0 \\ 0 & 0 & 0 \end{bmatrix} \quad (4)$$

where the rank of the covariance matrix $\Sigma_{\mathbf{x}\mathbf{x}}$ is two, i.e. it is singular and not invertible. Subsequent spherical normalization yields the final and general observation. We will refer to them as bearing vectors:

$$\mathbf{v} = \begin{bmatrix} v_x \\ v_y \\ v_z \end{bmatrix} = \frac{\mathbf{x}}{\|\mathbf{x}\|}, \Sigma_{\mathbf{v}\mathbf{v}} = \begin{bmatrix} \sigma_{v_x}^2 & \sigma_{v_{xy}} & \sigma_{v_{xz}} \\ \sigma_{v_{yx}} & \sigma_{v_y}^2 & \sigma_{v_{yz}} \\ \sigma_{v_{zx}} & \sigma_{v_{zy}} & \sigma_{v_z}^2 \end{bmatrix} \quad (5)$$

following (Förstner, 2010) the covariance is propagated using:

$$\Sigma_{\mathbf{v}\mathbf{v}} = \mathbf{J} \Sigma_{\mathbf{x}\mathbf{x}} \mathbf{J}^T, \mathbf{J} = \frac{1}{\|\mathbf{x}\|} (\mathbf{I}_3 - \mathbf{v}\mathbf{v}^T) \quad (6)$$

Observe, that the covariance matrix $\Sigma_{\mathbf{v}\mathbf{v}}$ remains singular, i.e. a Maximum Likelihood (ML) estimation based on the three residual components of bearing vectors is invalid. Thus, a minimal representation of the covariance information for the redundant representation of the homogeneous vector \mathbf{v} is desirable. In the following section, we will introduce the nullspace of vectors and show how this can be subsequently used to get an initial estimate

of the absolute orientation of the camera.

3.2 Nullspace of bearing vectors

The following was developed by (Förstner, 2010). The nullspace of \mathbf{v} , spans a two dimensional coordinate system whose axis, denoted as \mathbf{r} and \mathbf{s} , are perpendicular to \mathbf{v} and lie in its tangent space:

$$\mathbf{J}_{\mathbf{v}_r}(\mathbf{v}) = \text{null}(\mathbf{v}^T) = [\mathbf{r} \quad \mathbf{s}] = \begin{bmatrix} r_1 & s_1 \\ r_2 & s_2 \\ r_3 & s_3 \end{bmatrix} \quad (7)$$

The function $\text{null}(\cdot)$ calculates the Singular Value Decomposition (SVD) of \mathbf{v} and takes the two eigenvectors corresponding to the two zero eigenvalues. Further we assume $\mathbf{J}_{\mathbf{v}_r}$ to be an orthonormal matrix, i.e. $\mathbf{J}_{\mathbf{v}_r}^T(\mathbf{v})\mathbf{J}_{\mathbf{v}_r}(\mathbf{v}) = \mathbf{I}_2$. Note, that $\mathbf{J}_{\mathbf{v}_r}$ in addition represents the Jacobian of the transformation from the tangent space to the original vector. Thus the transpose $\mathbf{J}_{\mathbf{v}_r}^T$ yields the transformation from the original homogeneous vector \mathbf{v} to its reduced equivalent \mathbf{v}_r .

$$\mathbf{v}_r = \begin{bmatrix} dr \\ ds \end{bmatrix} = \mathbf{J}_{\mathbf{v}_r}^T(\mathbf{v})\mathbf{v} = \mathbf{0} \quad (8)$$

with nonsingular covariance

$$\Sigma_{\mathbf{v}_r\mathbf{v}_r} = \mathbf{J}_{\mathbf{v}_r}^T(\mathbf{v})\Sigma_{\mathbf{v}\mathbf{v}}\mathbf{J}_{\mathbf{v}_r}(\mathbf{v}) = \begin{bmatrix} \sigma_{v_{rx}}^2 & \sigma_{v_{ry}} \\ \sigma_{v_{rx}} & \sigma_{v_{ry}}^2 \end{bmatrix} \quad (9)$$

Another way to think about \mathbf{v}_r is as a residual in the tangent space. In the following, we will exploit Eq. 8 to get a linear estimate of the rotation and translation of the camera in the world frame by minimizing this residual in the tangent space.

3.3 Linear estimation of the camera pose

Using Eq. 1 and 7 we can reformulate Eq. 8:

$$\begin{bmatrix} dr \\ ds \end{bmatrix} = \begin{bmatrix} \mathbf{r}^T \\ \mathbf{s}^T \end{bmatrix} \lambda_i^{-1}(\mathbf{R}\mathbf{p}_i + \mathbf{t}) = \mathbf{0} \quad (10)$$

with $\lambda_i \neq 0$. Thus, if we knew the absolute orientation of our camera, the projection of a world point p_i to the tangent space of \mathbf{v} , should result in the same reduced coordinates, i.e. zero residual. Expanding Eq. 10 yields:

$$\begin{aligned} 0 &= r_1(\hat{r}_{11}p_x + \hat{r}_{12}p_y + \hat{r}_{13}p_z + \hat{t}_1) \\ &\quad + r_2(\hat{r}_{21}p_x + \hat{r}_{22}p_y + \hat{r}_{23}p_z + \hat{t}_2) \\ &\quad + r_3(\hat{r}_{31}p_x + \hat{r}_{32}p_y + \hat{r}_{33}p_z + \hat{t}_3) \\ 0 &= s_1(\hat{r}_{11}p_x + \hat{r}_{12}p_y + \hat{r}_{13}p_z + \hat{t}_1) \\ &\quad + s_2(\hat{r}_{21}p_x + \hat{r}_{22}p_y + \hat{r}_{23}p_z + \hat{t}_2) \\ &\quad + s_3(\hat{r}_{31}p_x + \hat{r}_{32}p_y + \hat{r}_{33}p_z + \hat{t}_3) \end{aligned} \quad (11)$$

with $\mathbf{p} = [p_x, p_y, p_z]^T$. Now, both equations are linear in the unknowns, i.e. we can stack them in a design matrix \mathbf{A} to obtain a homogeneous system of linear equations:

$$\mathbf{Au} = \mathbf{0} \quad (12)$$

with $\mathbf{u} = [\hat{r}_{11}, \hat{r}_{12}, \hat{r}_{13}, \hat{r}_{21}, \hat{r}_{22}, \hat{r}_{23}, \hat{r}_{31}, \hat{r}_{32}, \hat{r}_{33}, \hat{t}_1, \hat{t}_2, \hat{t}_3]^T$. As each observation yields two residuals, at least $I > 5$ points are necessary to solve Eq. 12. Assuming uncorrelated observa-

tions, the stochastic model is given by:

$$\mathbf{P} = \begin{bmatrix} \Sigma_{\mathbf{v}_r^1 \mathbf{v}_r^1}^{-1} & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & \Sigma_{\mathbf{v}_r^i \mathbf{v}_r^i}^{-1} \end{bmatrix} \quad (13)$$

and the final normal equations are:

$$\mathbf{A}^T \mathbf{P} \mathbf{A} \mathbf{u} = \mathbf{N} \mathbf{u} = \mathbf{0} \quad (14)$$

We find the \mathbf{u} that minimizes Eq. 14 subject to $\|\mathbf{u}\| = 1$, using SVD:

$$\mathbf{N} = \mathbf{U} \mathbf{D} \mathbf{V}^T \quad (15)$$

The solution is the particular column of \mathbf{V} that corresponds to the smallest singular value in \mathbf{D} :

$$\hat{\mathbf{R}} = \begin{bmatrix} \hat{r}_{11} & \hat{r}_{12} & \hat{r}_{13} \\ \hat{r}_{21} & \hat{r}_{22} & \hat{r}_{23} \\ \hat{r}_{31} & \hat{r}_{32} & \hat{r}_{33} \end{bmatrix}, \mathbf{t} = [\hat{t}_1, \hat{t}_2, \hat{t}_3] \quad (16)$$

It is determined up to a scale factor, thus the translational part $\hat{\mathbf{t}}$ only points in the right direction. The scale can be recovered from the fact, that the norm of each column $\hat{\mathbf{r}}_1$, $\hat{\mathbf{r}}_2$ and $\hat{\mathbf{r}}_3$ of the rotation matrix $\hat{\mathbf{R}}$ must equal one. Hence, the final translation is:

$$\mathbf{t} = \frac{\hat{\mathbf{t}}}{\sqrt[3]{\|\hat{\mathbf{r}}_1\| \|\hat{\mathbf{r}}_2\| \|\hat{\mathbf{r}}_3\|}} \quad (17)$$

The exploited constrain shows, that the 9 rotation parameters do not define a correct rotation matrix. This, can be solved by calculating the SVD of $\hat{\mathbf{R}}$:

$$\hat{\mathbf{R}} = \mathbf{U}_R \mathbf{D}_R \mathbf{V}_R^T \quad (18)$$

and the best rotation matrix minimizing the Frobenius norm is found as:

$$\mathbf{R} = \mathbf{U}_R \mathbf{V}_R^T. \quad (19)$$

Up to this point, a linear ML estimation of the absolute camera pose is obtained. To increase the accuracy, a non-linear refinement procedure is used. Doing a subsequent refinement of the initial estimate is a common procedure, e.g. performed in (Lepetit et al., 2009), (Ferraz et al., 2014b) or (Lu et al., 2000).

3.4 Non-linear refinement

We apply a Gauss-Newton optimization to iteratively refine the camera pose. Specifically, we minimize the tangent space residuals defined in Eq. 10. This is reasonably fast for two reasons. On the one hand, the nullspace vectors are already calculated and we simply have to calculate the dot products between the tangent space vectors and each transformed world point. On the other hand, the results of the linear estimates are already close to a local minimum, i.e. the Gauss-Newton optimization converges quickly. In practice we found, that a maximum number of five iterations is sufficient. To arrive at a minimal representation of the rotation matrix, we chose to express \mathbf{R} in terms of the Cayley parametrization (Cayley, 1846).

3.5 Planar case

In the planar case, the SVD (Eq.15) yields up to four solution vectors, as the corresponding singular values become small (close to zero). In this case, the solution is a linear combination of those vectors. To solve for the coefficients, an equation system with non-linear constraints had to be solved. We avoid this using the following trick. Let $\mathbf{M} = [\mathbf{p}_1, \mathbf{p}_2, \dots, \mathbf{p}_i]$ be a $3 \times I$ matrix of all

world points. The eigenvalues of $\mathbf{S} = \mathbf{M} \mathbf{M}^T$ give us information about the distribution of the world points. In the ordinary 3D case, the rank of matrix \mathbf{S} is three and the smallest eigenvalue is not close to zero. In the planar case, the smallest eigenvalue becomes small and the rank of matrix \mathbf{S} is two. If the world points lie on a plane that is spanned by two coordinate axis, respectively, i.e. one of the elements of all world points is a constant, we could simply omit the corresponding column from matrix \mathbf{A} and get a distinct solution.

In general, the points can lie on an arbitrary plane in the world frame. Thus, we use the eigenvectors of \mathbf{S} as a rotation matrix \mathbf{R}_S and rotate the world points to a new frame using:

$$\hat{\mathbf{p}}_i = \mathbf{R}_S^T \mathbf{p}_i \quad (20)$$

Here, we can identify the constant element of the coordinates and omit the corresponding column from the design matrix \mathbf{A} . Note, that this transformation does not change the structure of the problem. The rotation matrix obtained after SVD (Eq. 19), simply has to be rotated back to the original coordinate frame:

$$\mathbf{R} = \mathbf{R}_S \mathbf{R} \quad (21)$$

To keep our method as general as possible, the matrix \mathbf{S} is always calculated. We then simply switch between the planar and the ordinary case, depending on the rank of the matrix \mathbf{S} . To determine the rank we use rank-revealing QR decomposition with full pivoting and set a threshold (1e-10) on the smallest eigenvalue.

4. RESULTS

In this section, we compare our algorithm to all state-of-the-art algorithms using synthetic as well as real data. To reasonably assess the runtime performance of each algorithm, we categorize the state-of-the-art solvers according to their implementation:

Matlab

OPnP (Zheng et al., 2013a), EPPnP (Ferraz et al., 2014b), CEPPnP (Ferraz et al., 2014a), DLS (Hesch and Roumeliotis, 2011), LHM (Lu et al., 2000), ASPnP (Zheng et al., 2013b), SDP (Schweighofer and Pinz, 2008), RPnP (Li et al., 2012), PPnP (Garro et al., 2012)

C++,mex

UPnP (Kneip et al., 2014), EPnP+GN (Lepetit et al., 2009), MLPnP (this paper)

Note however, that the Matlab implementations of many algorithms are already quite optimized, and it is unclear how large the performance increase of a C++ version would be.

Our method is implemented in both Matlab and C++. We integrated the C++ version in OpenGV (Kneip and Furgale, 2014). The Matlab version will be made publicly available from the website of the corresponding author, i.e. all results are reproducible. All experiments have been conducted on a Laptop with Intel Core i7-3630QM@2.4Ghz.

4.1 Synthetic experiments

We use the Matlab toolbox provided by the authors of (Ferraz et al., 2014a, Li et al., 2012, Zheng et al., 2013a) and compare our algorithm in terms of accuracy and speed. The simulation configurations as well as the evaluation metrics are also part of the

toolbox and are briefly described in the following. All experiments are repeated $T = 250$ times and the mean pose errors are reported.

Assume a virtual calibrated camera with a focal length of $f = 800$ pixels. First, 3D points are randomly sampled in the camera frame on the interval $[-2, 2] \times [-2, 2] \times [4, 8]$ and projected to the image plane. Then, the image plane coordinates \mathbf{x}' are perturbed by Gaussian noise with different standard deviations σ yielding a covariance matrix $\Sigma_{\mathbf{x}'\mathbf{x}'}$ for each feature. At this point, we arrived at Eq. 2.

Now, the image plane coordinates \mathbf{x}' are usually transformed to normalized image plane coordinates applying the forward projection π . In the synthetic experiment, this simplifies to perspective division: $\mathbf{x} = [x'/f, y'/f, 1]^T$. Most algorithms work with the first two elements of \mathbf{x} , i.e. normalized image plane coordinates. Instead, we apply Eq. 5 and spherically normalize the vectors, to simulate a general camera model. Further, we propagate the covariance information using Eq. 6.

Finally, the ground truth translation vector \mathbf{t}_{gt} is chosen as the centroid of the 3D points and the ground truth rotation \mathbf{R}_{gt} is randomly sampled. Subsequently, all points are transformed to the world frame yielding \mathbf{p}_i .

The rotation accuracy in degree between \mathbf{R}_{gt} and \mathbf{R} is measured as $\max_{k=1}^3(\arccos(\mathbf{r}_{k,gt}^T \cdot \mathbf{r}_k) \times 180/\pi)$, where $\mathbf{r}_{k,gt}$ and \mathbf{r} are the k-th column of the respective rotation matrix. The translation accuracy is computed in % as $\|\mathbf{t}_{gt} - \mathbf{t}\|/\|\mathbf{t}\| \times 100$.

The first column of 3 depicts the ordinary and the second column the planar case, where the Z coordinate of the world points is set to zero. The first two rows of Fig. 3 depict the mean accuracy for rotation and translation for an increasing number of points ($I=10, \dots, 200$). Following the experiments of (Ferraz et al., 2014a), the image plane coordinates are perturbed by Gaussians with $\sigma = [1, \dots, 10]$ and 10% of the features are perturbed by each noise level, respectively.

For the third and fourth row of Fig. 3 the number of features is kept constant ($I=50$) and only the noise level is increased. This time, a random σ is chosen for each feature from 0 to the maximum number, ranging from 0 to 10.

The experiments for the ordinary case show, that MLPnP outperforms all other state-of-the-art algorithms in terms of accuracy. Moreover, it stays among the fastest algorithms as depicted in Fig. 2. For less than 20 points, the algorithm is even faster than EPnP, that is still the fastest PnP solution.

4.2 Real Data

To evaluate the performance of our proposed MLPnP method on real data, we recorded a trajectory of $T = 200$ poses of a moving, head mounted fisheye camera in an indoor environment. The camera is part of a helmet multi-camera system depicted in Fig. 4a. We use the camera model of (Scaramuzza et al., 2006) and the toolbox of (Urban et al., 2015) to calibrate the camera. All sensors have a resolution of 754×480 pixels and are equipped with equal fisheye lenses. The field of view of each lens covers 185° . The ground truth poses \mathbf{M}_{gt}^t are obtained by a Leica T-Probe, mounted onto the helmet and tracked by a Leica lasertracker. This tracker was in addition used to measure the 3D coordinates of the world points. The position accuracy of this system is $\sigma_{pos} \approx 100\mu m = 0.1mm$ and the angle accuracy is indicated with $\sigma_{angle} \approx 0.05mrad$.

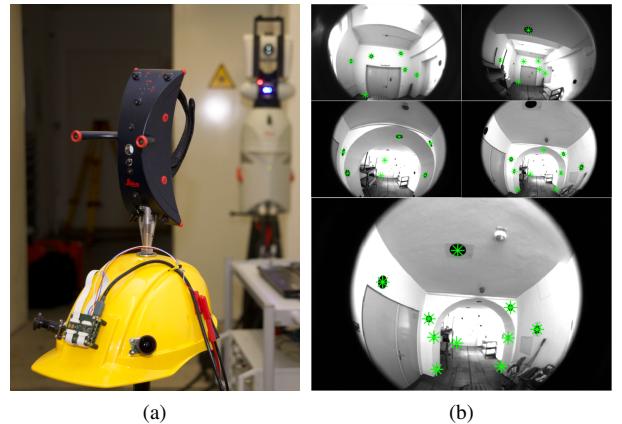


Figure 4: (a) Camera system with mounted T-Probe. The background shows the lasertracker. (b) Some images taken from the trajectory. Tracked world points are depicted in green.

For each frame t in the trajectory, the epicenters of I passpoints (planar, black filled circles) are tracked over time, as depicted in Fig. 4b. The number of passpoints I , that are visible in each frame ranges from 6 to 13.

To assess the quality of the estimated camera pose, we calculate the mean standard deviation (SD) $\bar{\sigma}_{\mathbf{R}}$ and $\bar{\sigma}_{\mathbf{t}}$ of the relative orientation between camera and T-Probe. The relative orientation for each frame is calculated as follows:

$$\mathbf{M}_{rel} = \mathbf{M}_{gt}^{-1} \mathbf{M}_t \quad (22)$$

where $\mathbf{M}_t = [\mathbf{R}, \mathbf{t}]$ is the current camera pose and is estimated by each of the 13 algorithms respectively.

4.3 Including Covariance Information

For the first frame $t = 1$, the covariance information about the measurements $\Sigma_{\mathbf{x}'\mathbf{x}'}$ is set to the identity \mathbf{I}_2 for each feature. To this point, the only thing we know about the features is, that they are measured with identical accuracy of 1 pixel. For the following frame $t = 2$, however, estimation theory can tell us, how well a point was measured in the previous frame, given the estimated parameter vector.

Let \mathbf{A} be the Jacobian of our residual function Eq. 10. Then, the covariance matrix of the unknown rotation and translation parameters is:

$$\Sigma_{\hat{\mathbf{r}}\hat{\mathbf{t}}} = (\mathbf{A}^T \mathbf{P} \mathbf{A})^{-1} \quad (23)$$

and cofactor matrix of the observations:

$$\mathbf{Q}_{\mathbf{v}_r \mathbf{v}_r} = \mathbf{A} \Sigma_{\hat{\mathbf{r}}\hat{\mathbf{t}}} \mathbf{A}^T \quad (24)$$

Observe, that this gives us the cofactor matrix in the reduced observation space. Thus, we project them back to the true observation space of the respective point i :

$$\Sigma_{\mathbf{v}_i \mathbf{v}_i} = \mathbf{J}_{\mathbf{v}_r^i} (\mathbf{v}_i) \mathbf{Q}_{\mathbf{v}_r^i \mathbf{v}_r^i} \mathbf{J}_{\mathbf{v}_r^i}^T (\mathbf{v}_i) \quad (25)$$

Now, we can set the covariance matrices $\Sigma_{\mathbf{v}_i \mathbf{v}_i}$ for frame $t = 2$ to the estimated values from the first frame $t = 1$ and so on. The results from this method are depicted as MLPnP+ Σ in Fig. 5. Clearly, the consistent use of the full covariance information gives another performance increase. Fig. 5a shows the mean standard deviation of the three orientation components.

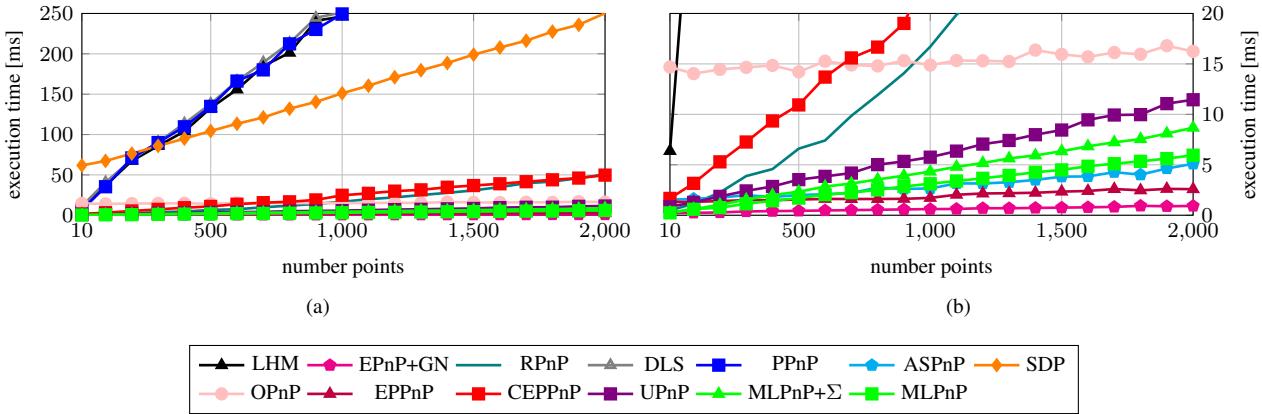


Figure 2: Runtime. (a) All methods. (b) Zoom to the fastest methods. CEPPnP and MLPnP+ Σ incorporate measurement uncertainty. All other algorithms assume equally well measured image points.

4.4 External vs. internal accuracies

Usually, it is desirable to have a measure of reliability about the result of an algorithm. A common approach in geometric vision, is to quantify the quality of the pose in terms of the reprojection error. Most of the time, this might be sufficient. Many computer vision systems are based on geometric algorithms and need a measure of how accurate and robust objects are mapped in the camera images. But sometimes it is useful to get more information about the quality of the estimated camera pose, e.g. for probabilistic SLAM systems. As MLPnP is a full ML estimator, we can employ estimation theory to obtain the covariance information about the estimated pose parameters and thus their standard deviations as a measure of reliability. Let \mathbf{r} be the vector of stacked residuals $[dr_i, ds_i]^T$. First, we calculate the variance factor:

$$\sigma_0^2 = \frac{\mathbf{r}^T \mathbf{P} \mathbf{r}}{b} \quad (26)$$

with \mathbf{P} being the stochastic model and $b = 2I - 6$ is the redundancy of our problem. Then, the 6×6 covariance matrix $\Sigma_{\hat{\mathbf{r}}\hat{\mathbf{t}}}$ of the unknowns is extracted with Eq. 23. Finally, the 6×1 vector of standard deviations of the camera pose is:

$$\sigma_{\hat{\mathbf{r}}, \hat{\mathbf{t}}} = \sigma_0 \sqrt{\text{diag}(\Sigma_{\hat{\mathbf{r}}\hat{\mathbf{t}}})} \quad (27)$$

We extract the covariance matrix for every frame in the trajectory and calculate the mean standard deviation over all frames. Tab. 2 depicts the results compared to the standard deviation that was calculated to the ground truth. The marginal differences between the internal uncertainty estimations and the external ground truth uncertainties is an empirical proof that our proposed ML estimator is statistically optimal.

| | σ_{angle} [deg] | σ_{pos} [cm] |
|--------------|-------------------------------|----------------------------|
| ground truth | 0.19 | 1.03 |
| estimated | 0.18 | 0.92 |

Table 2: Comparison between the estimated (internal) uncertainty and the uncertainty obtained by the ground truth (external)

5. CONCLUSION

In this paper, we proposed a Maximum Likelihood solution to the PnP problem. First, the variance propagation from 2D image

observations to general 3D bearing vectors was introduced. The singularity of the resulting 3×3 covariance matrices of such bearing vectors motivated the subsequent reduction of the covariance to the vector tangent space. In addition, this reduced formulation allows to obtain a solution to the PnP problem in terms of a linear Maximum Likelihood estimation followed by a fast iterative Gauss-Newton refinement. Finally the method was tested and evaluated against all state-of-the-art PnP methods. It shows similar execution times compared to the fastest methods and outperforms the state-of-the-art in terms of accuracy.

ACKNOWLEDGMENT

This project was partially funded by the DFG research group FG 1546 "Computer-Aided Collaborative Subway Track Planning in Multi-Scale 3D City and Building Models". In addition the authors would like to thank Patrick Erik Bradley for helpful discussions and insights.

REFERENCES

- Ansar, A. and Daniilidis, K., 2003. Linear pose estimation from points or lines. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)* 25(5), pp. 578–589.
- Cayley, A., 1846. About the algebraic structure of the orthogonal group and the other classical groups in a field of characteristic zero or a prime characteristic. *Reine Angewandte Mathematik*.
- Choi, C. and Christensen, H. I., 2012. Robust 3D visual tracking using particle filtering on the special euclidean group: A combined approach of keypoint and edge features. *International Journal of Robotics Research (IJRR)* 31(4), pp. 498–519.
- Collet, A., Berenson, D., Srinivasa, S. S. and Ferguson, D., 2009. Object recognition and full pose registration from a single image for robotic manipulation. In: *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)*, pp. 48–55.
- DeMenthon, D. and Davis, L. S., 1992. Exact and approximate solutions of the perspective-three-point problem. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)* 14(11), pp. 1100–1105.
- Ferraz, L., Binefa, X. and Moreno-Noguer, F., 2014a. Leveraging feature uncertainty in the PnP problem. In: *Proceedings of the British Machine Vision Conference (BMVC)*.
- Ferraz, L., Binefa, X. and Moreno-Noguer, F., 2014b. Very fast solution to the PnP problem with algebraic outlier rejection. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 501–508.

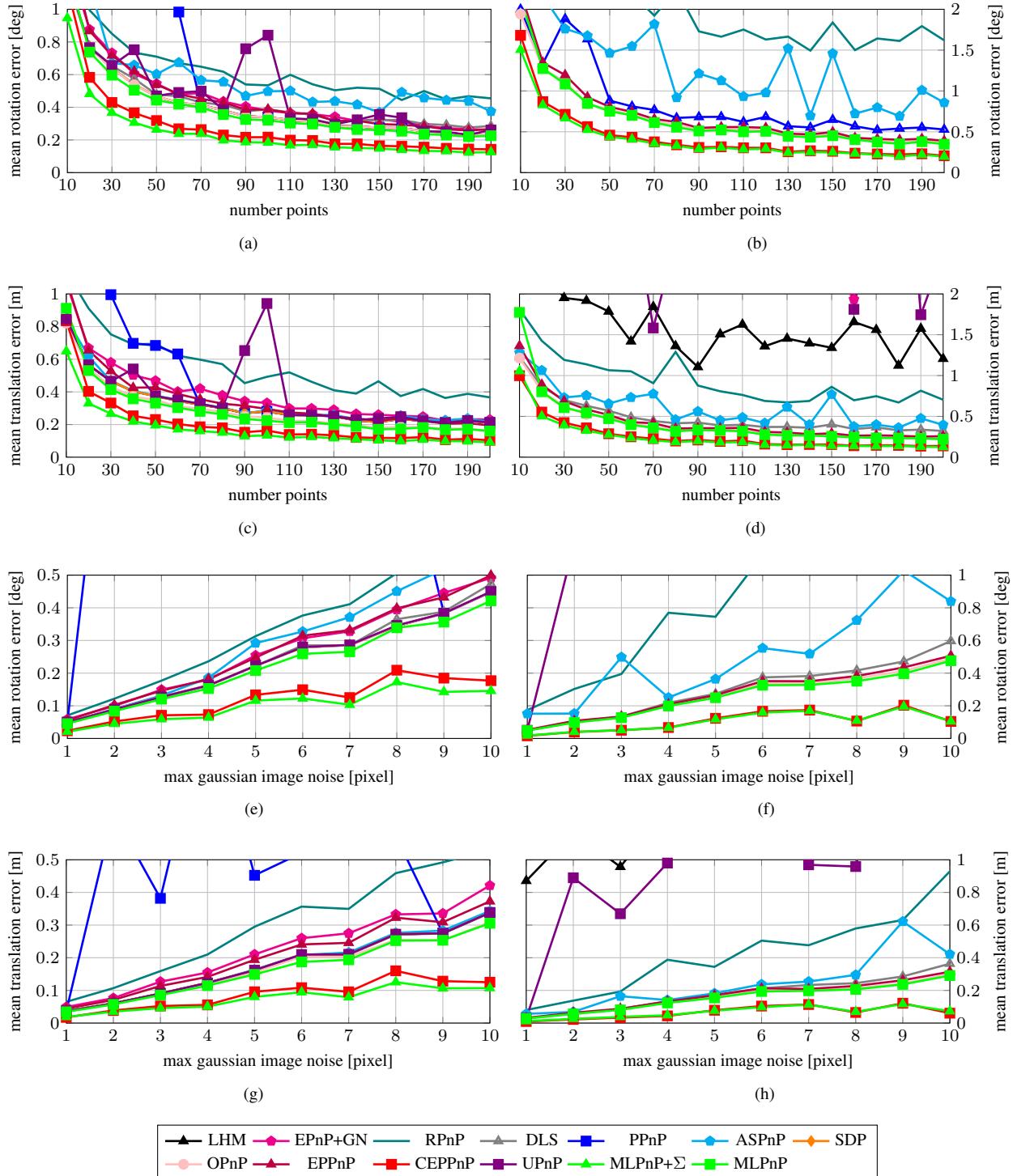


Figure 3: Left column ordinary 3D. Right column planar case. (a)-(d) increasing number of points. (e)-(h) 50 points fix, increasing noise level. CEPPnP and MLPnP+ Σ incorporate measurement uncertainty. All other algorithms assume that all points are measured with identical accuracy.

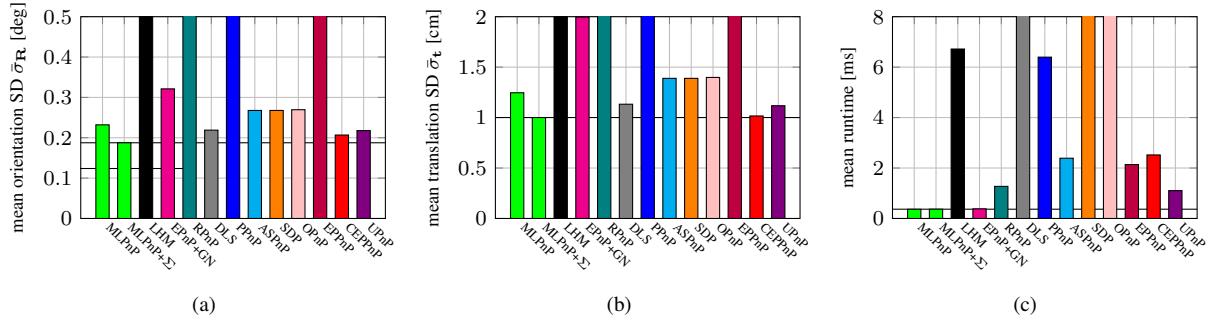


Figure 5: (a) mean orientation and (b) mean translation standard deviation over all frames of the real trajectory. (c) mean runtime over all frames. CEPPnP and MLPnP+ Σ incorporate measurement uncertainty. All other algorithms assume equally well measured image points.

- Fiore, P. D., 2001. Efficient linear solution of exterior orientation. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)* 23(2), pp. 140–148.
- Fischler, M. A. and Bolles, R. C., 1981. Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. *Communications of the ACM* 24(6), pp. 381–395.
- Förstner, W., 2010. Minimal representations for uncertainty and estimation in projective spaces. In: *Proceedings of the Asian Conference on Computer Vision (ACCV)*, Springer, pp. 619–632.
- Garro, V., Crosilla, F. and Fusielo, A., 2012. Solving the PnP problem with anisotropic orthogonal procrustes analysis. In: *Second International Conference on 3D Imaging, Modeling, Processing, Visualization & Transmission*, pp. 262–269.
- Haralick, R. M., Lee, D., Ottenburg, K. and Nolle, M., 1991. Analysis and solutions of the three point perspective pose estimation problem. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 592–598.
- Hartley, R. and Zisserman, A., 2003. *Multiple view geometry in computer vision*. Cambridge University Press, New York, NY, USA.
- Hesch, J. A. and Roumeliotis, S. I., 2011. A direct least-squares (DLS) method for PnP. In: *Proceedings of the International Conference on Computer Vision (ICCV)*, pp. 383–390.
- Horraud, R., Conio, B., Leboulleux, O. and Lacolle, L. B., 1989. An analytic solution for the perspective 4-point problem. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 500–507.
- Hu, Z. and Wu, F., 2002. A note on the number of solutions of the noncoplanar P4P problem. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)* 24(4), pp. 550–555.
- Kneip, L. and Furgale, P., 2014. OpenGV: A unified and generalized approach to real-time calibrated geometric vision. In: *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)*, pp. 1–8.
- Kneip, L., Li, H. and Seo, Y., 2014. UPnP: An optimal O(n) solution to the absolute pose problem with universal applicability. In: *Proceedings of the European Conference on Computer Vision (ECCV)*, Springer, pp. 127–142.
- Kneip, L., Scaramuzza, D. and Siegwart, R., 2011. A novel parametrization of the perspective-three-point problem for a direct computation of absolute camera position and orientation. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2969–2976.
- Kukelova, Z., Bujnak, M. and Pajdla, T., 2013. Real-time solution to the absolute pose problem with unknown radial distortion and focal length. In: *Proceedings of the International Conference on Computer Vision (ICCV)*, pp. 2816–2823.
- Lepetit, V., Moreno-Noguer, F. and Fua, P., 2009. EPnP: An accurate O(n) solution to the PnP problem. *International Journal of Computer Vision (IJCV)* 81(2), pp. 155–166.
- Li, S. and Xu, C., 2011. A stable direct solution of perspective-three-point problem. *International Journal of Pattern Recognition in Artificial Intelligence* 25(05), pp. 627–642.
- Li, S., Xu, C. and Xie, M., 2012. A robust O(n) solution to the perspective-n-point problem. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)* 34(7), pp. 1444–1450.
- Lu, C.-P., Hager, G. D. and Mjolsness, E., 2000. Fast and globally convergent pose estimation from video images. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)* 22(6), pp. 610–622.
- Luhmann, T., Robson, S., Kyle, S. and Harley, I., 2006. *Close range photogrammetry: Principles, methods and applications*. Whittles Publishing, Dunbeath, Scotland.
- Moreno-Noguer, F., Lepetit, V. and Fua, P., 2007. Accurate non-iterative O(n) solution to the PnP problem. In: *Proceedings of the International Conference on Computer Vision (ICCV)*, pp. 1–8.
- Müller, M., Rassweiler, M.-C., Klein, J., Seitl, A., Gondan, M., Baumhauer, M., Teber, D., Rassweiler, J. J., Meinzer, H.-P. and Maier-Hein, L., 2013. Mobile augmented reality for computer-assisted percutaneous nephrolithotomy. *International journal of computer assisted radiology and surgery* 8(4), pp. 663–675.
- Quan, L. and Lan, Z., 1999. Linear n-point camera pose determination. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)* 21(8), pp. 774–780.
- Scaramuzza, D., Martinelli, A. and Siegwart, R., 2006. A flexible technique for accurate omnidirectional camera calibration and structure from motion. In: *Proceedings of the Fourth IEEE International Conference on Computer Vision Systems (ICVS)*, pp. 45–45.
- Schweighofer, G. and Pinz, A., 2008. Globally optimal O(n) solution to the PnP problem for general camera models. In: *Proceedings of the British Machine Vision Conference (BMVC)*, pp. 1–10.
- Triggs, B., 1999. Camera pose and calibration from 4 or 5 known 3D points. In: *Proceedings of the International Conference on Computer Vision (ICCV)*, Vol. 1, pp. 278–284.
- Urban, S., Leitloff, J. and Hinz, S., 2015. Improved wide-angle, fisheye and omnidirectional camera calibration. *ISPRS Journal of Photogrammetry and Remote Sensing* 108, pp. 72–79.
- Weinmann, M., Weinmann, M., Hinz, S. and Jutzi, B., 2011. Fast and automatic image-based registration of TLS data. *ISPRS Journal of Photogrammetry and Remote Sensing* 66(6), pp. S62–S70.
- Wu, C., 2015. P3.5P: Pose estimation with unknown focal length. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2440–2448.
- Yang, L., Wang, J., Ando, T., Kubota, A., Yamashita, H., Sakuma, I., Chiba, T. and Kobayashi, E., 2015. Vision-based endoscope tracking for 3D ultrasound image-guided surgical navigation. *Computerized Medical Imaging and Graphics* 40, pp. 205–216.
- Zheng, Y., Kuang, Y., Sugimoto, S., Astrom, K. and Okutomi, M., 2013a. Revisiting the PnP problem: A fast, general and optimal solution. In: *Proceedings of the International Conference on Computer Vision (ICCV)*, pp. 2344–2351.
- Zheng, Y., Sugimoto, S. and Okutomi, M., 2013b. ASPnP: An accurate and scalable solution to the perspective-n-point problem. *IEICE Transactions on Information and Systems* 96(7), pp. 1525–1535.