Insights into the robustness of control point configurations for homography and planar pose estimation

Raul Acuna[1] and Volker Willert[1]

***Abstract*— In this paper, we investigate the influence of the spatial configuration of a number of $n \geq 4$ control points on the accuracy and robustness of space resection methods, e.g. used by a fiducial marker for pose estimation. We find robust configurations of control points by minimizing the first order perturbed solution of the DLT algorithm which is equivalent to minimizing the condition number of the data matrix. An empirical statistical evaluation is presented verifying that these optimized control point configurations not only increase the performance of the DLT homography estimation but also improve the performance of planar pose estimation methods like IPPE and EPnP, including the iterative minimization of the reprojection error which is the most accurate algorithm. We provide the characteristics of stable control point configurations for real-world noisy camera data that are practically independent on the camera pose and form certain symmetric patterns dependent on the number of points. Finally, we present a comparison of optimized configuration versus the number of control points.**

I. INTRODUCTION

The Perspective-n-Point (PnP) problem and the special case of planar pose estimation via homography estimation are some of the most researched topics in the fields of computer vision and photogrammetry. Even though the research in these areas has been wide, there is a surprising lack of information regarding the effect of 3D control point configurations on the accuracy and robustness of the estimation methods.

As shown in Sec. II, it is clear from the literature, that control point configurations are relevant and they do influence the accuracy and robustness of pose estimates. However, the findings are rather general, since they are based on hands-on experience and thus far only lead to some rules of thumb. Most obvious and widely accepted is, that increasing the number of control points increases the accuracy of the results in presence of noise. Further on, in several studies when simulations are performed to compare methods, great care is given to possible singular point configurations, such as non-centered data or near-planar cases which are singularities or degenerate cases for certain estimation methods, so there is a need at least to find out which point configurations are better than others so fair comparisons can be made.

A more thorough evaluation is given for the *normalized* DLT algorithm, whereas the normalization has already shown to improve the estimation because it is related to the condition number of the set of DLT equations [1]. The only error analysis for homography estimation found so far by the authors in the literature presents a statistical analysis and simulations of the errors in the homography coefficients [2].

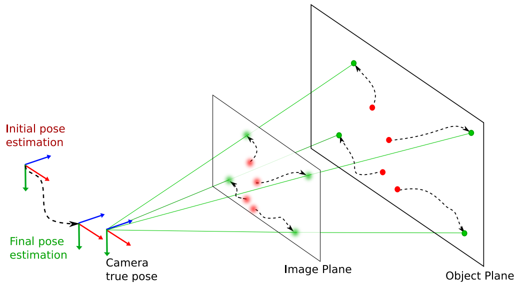

Fig. 1. Optimizing point configurations: Control points with known 3D coordinates on the object plane (marker) in arbitrary configuration (red) are moved towards optimized configurations (green) for pose estimation from these control points and their corresponding noisy projections on the image plane (blurry red and green). The control points' dynamics (11) is given by the gradient descent steps minimizing the optimization objective (10) that results in improved pose estimations (from red to green) close to the true camera pose (black).

However, none of the above give an answer to the question: *Are there optimized perspective-n-point configurations, which can increase the accuracy and robustness of space resection methods?*

If there are, this question includes several follow-up questions: Are the optimized configurations dependent on the pose, or is there only one configuration that is optimal for all poses? What are the specifics of this/these configuration(s) in relation to absolute coordinates and relative distances between coordinates? Are there similarities between configurations that differ in the number of control points? When does an increase in the number of points that are arbitrarily configured outperform the optimal configuration of a small number point set?

In this paper, we search for an answer to these questions in the planar case by proposing an optimization objective to find optimized planar control point configurations. Figure 1 sketches the main idea of optimizing the proposed objective via a gradient descent approach and the stepwise improvement of the accuracy of the pose estimate starting from some initial control point configuration. Each descent step leads to a change in control point configuration and thus defines a stable dynamics for the control points that are placed on a planar visual fiducial marker (object plane) converging to stable control point configurations.

The paper is structured as follows: In Sec. II, we classify pose estimation methods and summarize known findings

*This work was sponsored by the German Academic Exchange Service (DAAD) and the Becas Chile doctoral scholarship.

[1] These authors are within the Institute of Automatic Control and Mechatronics, Technische Universität Darmstadt, Germany. `(racuna, vwillert)@rmr.tu-darmstadt.de`

of control point configurations. In Sec. III, we derive the optimization objective based on golden standard algorithms for pose estimation. In Sec. V, we describe the simulation results, and finally, in Sec. VI, we discuss the results and give conclusions.

II. State of the art: brief history of space resection and PnP methods

Camera or space resection is a term used in the field of photogrammetry in which the spatial position and orientation of a photo are obtained by using image measurements of control points present on the photo, also know in the computer vision community as the Perspective-n-Point (PnP) problem. PnP can be considered an over-constrained (only for $n \geq 3$) and generic solution to the pose estimation problem from point correspondences. PnP methods can be classified into those which solve for a small and predefined number n of points, and those which can handle the general case. Several solutions have been presented in the literature [3], which in general provide four solutions for non-collinear points. Thus, prior knowledge has to be included to choose the correct solution.

Since it has been proven that pose accuracy usually increases with the number of points [3], other PnP approaches that use more points ($n > 3$) are usually preferred. The general PnP methods can be broadly divided into whether they are iterative or non-iterative. Iterative approaches formulate the problem as a non-linear least-squares problem. They differ in the choice of the cost function to minimize, which is usually associated to an algebraic or geometric error. Some of the most important iterative methods in chronological order are: the **POSIT** algorithm [4], the **LHM** [5], the Procrustes PnP method or **PPnP** [6] and the global optimization method **SDP** [7].

Most iterative methods have the disadvantage that they return only a single pose solution, which might not be the true one. Most of them can only guarantee a local minimum and the ones that find a global minimum remain computational intensive. The major limitation of iterative methods is that they are rather slow, neither convergence nor optimality can be guaranteed and a good initial guess is usually needed to converge to the right solution.

Non-iterative methods try to reformulate the problem so it may be solved by a potentially large equation system. However, early non-iterative solvers were also computational demanding and worse for a larger number of points. The first efficient and non-iterative $O(n)$ solution was **EPnP** [8], which was later improved by using an iterative method to increase accuracy.

More recent approaches are based on polynomial solvers trying to achieve linear performance without the problems of EPnP and with higher accuracy [9]–[13].

A special case of PnP is planar pose estimation, or PPE, which is a space resection problem that involves the process of recovering the relative pose of a plane with respect to a camera's coordinate frame from a single image measurement and which is the focus of this work. A PPE problem can be solved by calculating the object-plane to image-plane homography transformation and then extracting the pose from the homography matrix. This is known as homography decomposition [14], [15], or by using a set of points in the plane as the measurement with a special case of the PnP methods (planar PnP). Some of the most important planar PnP methods are the iterative **RPP-SP** [16] and the more recent direct method **IPPE** [17]. In general, planar PnP methods outperform the best homography decomposition methods when noise is present. Additionally, homography decomposition methods only provide a single solution in contrast to modern planar-PnP methods.

The standard linear algorithm for homography estimation is the Direct Linear Transform (DLT) [18], which was improved later in [19] using an orthogonalization step. For both methods, the normalization of the measurements is a key step to improve the quality of the estimated homography [18]. However, the normalization has some disadvantages [20]: First, the normalization matrices are calculated from noisy measurements and are sensitive to outliers, and second, for a given measurement the noise affecting each point is independent of the others.

A. Control points configurations

It has been pointed out in the literature [8], [10] that 3D point configurations have an influence on the local minima of the PnP problem. In the RPnP method paper [10], a broad classification of the control points configurations into three groups is presented. The classification is based on the rank of the matrix $\mathbf{M}^T\mathbf{M} \in \mathbb{R}^{3\times3}$, where $\mathbf{M} = [\mathbf{X}_1, \mathbf{X}_2, \dots, \mathbf{X}_n]^T$, $\mathbf{X}_i$ is the 3D coordinate of control point i and n is the amount of control points.

In EPnP [8] it is shown that if the control points are taken from the *uncentered data* or the region where the image projections of the control points cover only a small part of the image, the stability of the compared methods greatly degrades. In RPnP it is elaborated that based on the previous classification this *uncentered data* is a configuration that lays between the *ordinary 3D case* and the *planar case*.

Some assumptions about the influence of the control points configurations are also present in the IPPE paper [17]. Through statistical evaluations, the authors found out that the accuracy for the 4-point case decreases if the points are uniformly sampled from a given region. They circumvent this problem by selecting the corners of the region as the positions for the control points and then refer the reader to the Chen and Suter paper [2], where the analysis of the stability of the homography estimation to 1st order perturbations is presented. In this analysis, it is clear that the error in homography estimate is dependent on the singular values of the **A** matrix in the DLT algorithm (see also next section).

Additionally, in [21], [22] evaluations are presented characterizing pose-dependent offsets and uncertainty on the camera pose estimations. It is empirically proven by simulations that some poses of the camera are more stable for the estimation process than others.

III. Basics of golden standard algorithms for pose estimation

Before we explain the optimization method for optimizing point configurations, we shortly summarize the *golden standard* optimization methods for pose estimation from general and planar point configurations which are the minimization of the reprojection (geometric) error (MRE) for iterative methods and the minimization of the algebraic error for non-iterative methods via the DLT algorithm, respectively.

A. General point configuration for pose estimation

Given a 3D-2D point correspondence of i-th 3D control point p_i with world W coordinates $\mathbf{X}_i^W = [X_i^W, Y_i^W, Z_i^W]^T \in \mathbb{R}^3$ and its corresponding projection onto a planar calibrated camera[1] with normalized image coordinates $\mathbf{x}_i = [x_i, y_i]^T \in \mathbb{R}^2$ the relation between these points is given by the relative pose[2] $g = (\mathbf{R}, \mathbf{T})$ (Euclidean transformation) between world W and camera C frame $\mathbf{X}_i^C = \mathbf{R}\mathbf{X}_i^W + \mathbf{T}$ followed by a projection π with $\mathbf{x}_i = \pi(\mathbf{X}_i^C) = [X_i^C/Z_i^C, Y_i^C/Z_i^C]^T$.

This leads to the relation:

$$\mathbf{x}_i = \pi(\mathbf{X}_i^C) = \pi(\mathbf{R}\mathbf{X}_i^W + \mathbf{T})\,. \tag{1}$$

Including additive noise $\boldsymbol{\varepsilon}_i = [\varepsilon_i, \zeta_i]^T$ on the error-free image coordinates $\mathbf{x}_i$ we get noisy measurements of the image coordinates $\tilde{\mathbf{x}}_i = \mathbf{x}_i + \boldsymbol{\varepsilon}_i$. Thus, we can solve for the reprojection error $\|\boldsymbol{\varepsilon}_i\|_2^2 = \|\tilde{\mathbf{x}}_i - \mathbf{x}_i\|_2^2$ of each point which is a squared 2-norm. Minimizing the squared 2-norm of all points for the optimal pose $(\hat{\mathbf{R}}, \hat{\mathbf{T}})$ leads to the following least-squares estimator

$$(\hat{\mathbf{R}}, \hat{\mathbf{T}}) = \operatorname{argmin}_{\mathbf{R},\mathbf{T}} \sum_{i=1}^{n} \|\boldsymbol{\varepsilon}_i\|_2^2\,, \quad n \geq 3\,. \tag{2}$$

Iterative gradient descent optimization of (2) leads to the most accurate pose estimation results in the literature so far, also for planar point configurations.

B. Planar points configuration for pose estimation

If the control points $\mathbf{X}_i^W$ are all on a plane P, we can define a 2D subspace in the 3D world with coordinates[3] $\mathbf{X}_i^P = [X_i^P, Y_i^P]^T \in \mathbb{R}^2$. Plugging the planar constraint in equation (1), extending to homogeneous coordinates and rearranging the equation, leads to an homography mapping

$$\mathbf{X}_i^C = Z_i^C \overline{\mathbf{x}}_i = [\mathbf{r}_1, \mathbf{r}_2, \mathbf{T}]\overline{\mathbf{X}}_i^P = \mathbf{H}\overline{\mathbf{X}}_i^P\,. \tag{3}$$

Eliminating Z_i^C, we get $\overline{\mathbf{x}}_i \times \mathbf{H}\overline{\mathbf{X}}_i^P = 0$, where each point correspondence $\{\mathbf{x}_i, \mathbf{X}_i^P\}$ produces two linearly independent equations

$$\mathbf{A}_i\mathbf{h} = \begin{bmatrix} \mathbf{O}_{1\times 3} & -(\overline{\mathbf{X}}_i^P)^T & y_i(\overline{\mathbf{X}}_i^P)^T \\ (\overline{\mathbf{X}}_i^P)^T & \mathbf{O}_{1\times 3} & -x_i(\overline{\mathbf{X}}_i^P)^T \end{bmatrix} \begin{bmatrix} \mathbf{r}_1 \\ \mathbf{r}_2 \\ \mathbf{T} \end{bmatrix} = \mathbf{0}\,, \tag{4}$$

with $\mathbf{h} = [\mathbf{r}_1^T, \mathbf{r}_2^T, \mathbf{T}^T]^T \in \mathbb{R}^9$ and $\mathbf{A}_i \in \mathbb{R}^{2\times 9}$.

Again, assuming noisy measurements of the image coordinates $\tilde{\mathbf{x}}_i = \mathbf{x}_i + \boldsymbol{\varepsilon}_i$, we get noisy matrices

$$\begin{aligned} \tilde{\mathbf{A}}_i &= \begin{bmatrix} \mathbf{O}_{1\times 3} & -(\overline{\mathbf{X}}_i^P)^T & \tilde{y}_i(\overline{\mathbf{X}}_i^P)^T \\ (\overline{\mathbf{X}}_i^P)^T & \mathbf{O}_{1\times 3} & -\tilde{x}_i(\overline{\mathbf{X}}_i^P)^T \end{bmatrix} = \mathbf{A}_i + \mathbf{E}_i \qquad (5) \\ &= \begin{bmatrix} \mathbf{O}_{1\times 3} & -(\overline{\mathbf{X}}_i^P)^T & y_i(\overline{\mathbf{X}}_i^P)^T \\ (\overline{\mathbf{X}}_i^P)^T & \mathbf{O}_{1\times 3} & -x_i(\overline{\mathbf{X}}_i^P)^T \end{bmatrix} + \begin{bmatrix} \mathbf{O}_{1\times 6} & \zeta_i(\overline{\mathbf{X}}_i^P)^T \\ \mathbf{O}_{1\times 6} & \varepsilon_i(\overline{\mathbf{X}}_i^P)^T \end{bmatrix}. \end{aligned}$$

From $\tilde{\mathbf{A}}_i\mathbf{h} = (\mathbf{A}_i + \mathbf{E}_i)\mathbf{h}$ we can solve for the algebraic error $\|\mathbf{E}_i\mathbf{h}\|_2^2 = \|(\tilde{\mathbf{A}}_i - \mathbf{A}_i)\mathbf{h}\|_2^2 = \|\tilde{\mathbf{A}}_i\mathbf{h}\|_2^2$ of each point, because $\mathbf{A}_i\mathbf{h} = \mathbf{0}$ holds. Minimizing the squared 2-norm of all points for the optimal homography $\hat{\mathbf{h}}$ leads to the following least-squares estimator

$$\hat{\mathbf{h}} = \operatorname{argmin}_{\mathbf{h}} \sum_{i=1}^{n} \|\mathbf{E}_i\mathbf{h}\|_2^2\,, \quad s.t.\ \|\mathbf{h}\| = 1\,, \quad n \geq 4\,. \tag{6}$$

Since $\mathbf{h}$ contains 9 entries, but is defined only up to scale the total number of degrees of freedom is 8. Thus, the additional constraint $\|\mathbf{h}\| = 1$ is included to solve the optimization.

Now, stacking all $\{\tilde{\mathbf{A}}_i\}$ and $\{\mathbf{E}_i\}$ as $\tilde{\mathbf{A}} = [\tilde{\mathbf{A}}_1^T, \dots, \tilde{\mathbf{A}}_n^T]^T \in \mathbb{R}^{2n\times 9}$ and $\mathbf{E} = [\mathbf{E}_1^T, \dots, \mathbf{E}_n^T]^T \in \mathbb{R}^{2n\times 9}$ respectively, we arrive at solving the noisy homogeneous linear equation system

$$\tilde{\mathbf{A}}\mathbf{h} = \mathbf{E}\mathbf{h}\,. \tag{7}$$

The solution of (7) is equivalent to the solution of (6) and is given by the DLT algorithm applying a singular value decomposition (SVD) of $\tilde{\mathbf{A}} = \tilde{\mathbf{U}}\tilde{\mathbf{S}}\tilde{\mathbf{V}}^T$, whereas $\hat{\mathbf{h}} = \tilde{\mathbf{v}}_9$ with $\tilde{\mathbf{v}}_9$ being the right singular vector of $\tilde{\mathbf{A}}$, associated with the least singular value $\tilde{s}_9$. Usually, an additional normalization step of the coordinates of the control points and its projections is performed leading to the normalized DLT algorithm which is the golden standard for non-iterative pose estimation, because it is very easy to handle and serves as a basis for other non-iterative as well as iterative pose estimation methods.

IV. Optimizing points configuration for pose estimation

In order to find an optimal control points configurations, we need a proper optimization criterion. In the following, we propose an optimization criterion that is optimal for planar pose estimation using the (normalized) DLT algorithm, since it is the critical first step in planar pose estimation methods (even the gold standard of the minimization of the reprojection error requires a good initial guess, which is obtained from the DLT). We start with availing ourselves of perturbation theory applied to singular value decomposition of noisy matrices [23] and have a look at the first order perturbation expansion for the perturbed solution of the DLT algorithm, given in [2], which is

$$\hat{\mathbf{h}} = \tilde{\mathbf{v}}_9 = \mathbf{v}_9 - \sum_{k=1}^{8} \frac{\mathbf{u}_k^T\mathbf{E}\mathbf{v}_9}{s_k}\mathbf{v}_k\,. \tag{8}$$

Equation (8) clearly shows that the optimal solution for the homography that equals the right singular vector of the unperturbed matrix $\mathbf{A}$, associated with the least singular

[1] Assuming the calibration matrix $\mathbf{K} \in \mathbb{R}^{3\times 3}$ to be known, the homogeneous image coordinates in pixel $\overline{\mathbf{x}}_i' = [x_i', y_i', 1]^T$ can be transformed to homogeneous normalized image coordinates in metric units $\overline{\mathbf{x}}_i = \mathbf{K}^{-1}\overline{\mathbf{x}}_i'$.

[2] The rotation matrix is given by: $\mathbf{R} = [\mathbf{r}_1, \mathbf{r}_2, \mathbf{r}_3] \in \mathbb{R}^{3\times 3} | \mathbf{R}^T\mathbf{R} = \mathbf{I}, |\mathbf{R}| = 1$.

[3] Corresponding homogeneous coordinates are $\overline{\mathbf{X}}_i^P = [X_i^P, Y_i^P, 1]^T \in \mathbb{R}^3$.

value[4] $s_9 = 0$, is perturbed by the second term in (8). The second term is a weighted sum of the first eight optimal right singular vectors $\mathbf{v}_k$, whereas the weights $\mathbf{u}_k^T \mathbf{E} \mathbf{v}_9 / s_k$ are the influence of the measurement errors $\mathbf{E}$ on the unperturbed solution $\mathbf{v}_9$ along the different k dimensions of the model space. The presence of very small s_k in the denominator can give us very large weights for the corresponding model space basis vector $\mathbf{v}_k$ and dominate the error. Hence, small singular values s_k cause the estimation $\hat{\mathbf{h}}$ to be extremely sensitive to small amounts of noise in the data and correlates with the singular value spectrum[5] $(s_1 - s_8)$ as follows: The smaller the singular value spectrum, the less perturbed the estimation is. It is also well known, that the condition number of a matrix with respect to the 2-norm is given by the ratio between the largest and, in our case, second-smallest singular value [24]

$$c(\mathbf{A}) = \|\mathbf{A}\|_2 \|\mathbf{A}^{-1}\|_2 = \frac{s_{max}}{s_{min}} = \frac{s_1}{s_8}, \tag{9}$$

which is minimal if the singular value spectrum is minimal. The normalization of the control points and its projections which leads to the normalized DLT algorithm has already shown to improve the condition of matrix $\mathbf{A}$ [1]. Thus, we simply try to minimize the condition number c of matrix $\mathbf{A}$ with respect to all n control points $\{\mathbf{X}_i^P\}$ like follows:

$$\{\hat{\mathbf{X}}_i^P\} = \operatorname{argmin}_{\{\mathbf{X}_i^P\}} c\left(\mathbf{A}(\{\mathbf{X}_i^P\})\right). \tag{10}$$

Optimization of (10) is realized with a gradient descent minimization, whereas for calculation of the gradient vector we use automatic differentiation[6] [26]. This leads to the final discrete control points dynamics

$$\mathbf{X}_i^P(t+1) = \mathbf{X}_i^P(t) - \alpha(t) \nabla c\left(\mathbf{A}\left(\mathbf{X}_i^P(t)\right)\right), \tag{11}$$

for each iteration t and stepsize $\alpha(t)$, which is adapted using SuperSAB [27]. The control points dynamics can now be used to find optimal control point configurations for pose estimation from planar markers.

Given perturbations on the matrix $\mathbf{A}$ the relative error on the estimation of the homography parameters is defined as $\xi = (\mathbf{h} - \hat{\mathbf{h}})/\hat{\mathbf{h}}$ and from perturbation theory the following inequality defines an upper bound for the relative error:

$$\|\xi\| \leq c(\mathbf{A}) \|\mathbf{A} - \tilde{\mathbf{A}}\| / \|\mathbf{A}\|. \tag{12}$$

To find a lower bound, we can use the error of using a perturbed matrix $\tilde{\mathbf{A}}$ with the true homography $\mathbf{h}$ defined as $\tilde{\mathbf{A}}\mathbf{h}$ and the error of using the optimal homography estimation $\hat{\mathbf{h}}$ with the same perturbed matrix defined as $\tilde{\mathbf{A}}\hat{\mathbf{h}}$ to build the following inequality:

$$\|\tilde{\mathbf{A}}\mathbf{h} - \tilde{\mathbf{A}}\hat{\mathbf{h}}\| = \|\tilde{\mathbf{A}}(\mathbf{h} - \hat{\mathbf{h}})\| \leq \|\tilde{\mathbf{A}}\| \|\mathbf{h} - \hat{\mathbf{h}}\|, \tag{13}$$

which then divided by $\|\hat{\mathbf{h}}\|$ leads to a lower bound of the relative error:

$$\|\xi\| \geq \|\tilde{\mathbf{A}}(\mathbf{h} - \hat{\mathbf{h}})\| / (\|\tilde{\mathbf{A}}\| \|\hat{\mathbf{h}}\|). \tag{14}$$

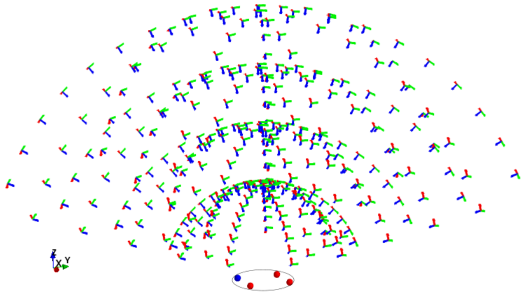

Fig. 2. Distribution of 400 camera poses used in simulations. A limiting circular plane including n control points is displayed at the bottom. The cameras are distributed evenly on spheres of evenly sampled radii, each one looking at the center of the circular plane keeping the complete circular plane in field of view.

The upper bound implies that there are only two ways to improve the maximum values of the relative error, either by reducing the perturbation of the $\mathbf{A}$ matrix, which usually can't be controlled, or by improving the condition number of the $\mathbf{A}$ matrix, which can be done by a normalization step in the DLT transform and by selecting optimized control point configurations.

V. Simulation and Real Experiment Results

Our simulation setup is based on a perspective camera model and a planar visual marker on $Z_i^W = 0$ centered in the origin $\mathbf{X}_o^W = [0,0,0]^T$ of world coordinates, we impose an arbitrary circular limit with a radius of $r = 0.15$ meters, this allows a smooth movement of the control points during the optimization while keeping them inside camera image. Rectangular limits were also tested but the discontinuities on the corners restrict the movement of the points.

A set of control points are randomly defined inside the limits of this circular plane, which are then projected onto the camera image[7]. A uniform distribution of 400 camera poses is defined around the marker as displayed in Fig. 2, this distribution provides a wide combination of rotation and translations (without lack of generalization) in the whole range of detection and allows us to properly compare the final point configurations in image coordinates.

A. Evaluations

To evaluate the improvement of the gradient descent optimization, we consider the optimization objective, which is the condition number (9) at each iteration t, given by $c(\mathbf{A}(t))$ in the DLT algorithm. To evaluate the effect of the optimization (10) on the underlying homography estimate $\hat{\mathbf{H}}(t)$ using a given set of n control points $\{\mathbf{X}_i^P\}(t)$, the movement of the points during the optimization is constrained to the limits of the circular bounds, we rely on the reprojection error $HE(\hat{\mathbf{H}}(t))$ induced by the estimated homography $\hat{\mathbf{H}}(t)$ given by

$$HE\left(\hat{\mathbf{H}}(t)\right) = \frac{1}{M} \sum_{j=1}^{M} \|\mathbf{x}_j(t) - \pi\left(\hat{\mathbf{H}}(t) \overline{\mathbf{X}}_j^P(t)\right)\|_2^2, \tag{15}$$

[4] The singular values are arranged in descending order: $s_1 \geq s_2 \geq \cdots \geq s_8 \geq s_9 = 0$.

[5] Here, the singular value spectrum between the first and second-last singular value is relevant, because $s_9 = 0$ holds.

[6] For implementation, we used *autograd* [25].

[7] Camera parameters: size $640 \times 480\,[pixel^2]$, intrinsic parameters $\mathbf{K} = [800, 0, 320; 0, 800, 240; 0, 0, 1]$.

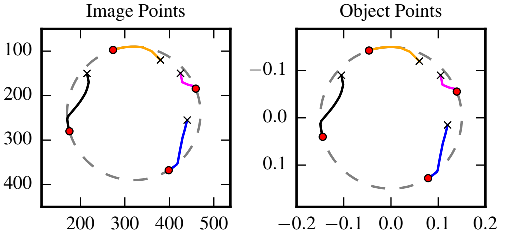

Fig. 3. (Fronto-Parallel). Movement of control points in image and object coordinates during optimization for a fronto-parallel camera configuration simulation until an optimized configuration (red dots) limited by the circle (dashed grey line).

for a fixed set of M validation control points $\{\mathbf{X}_j^P\} \notin \{\mathbf{X}_i^P\}(t)$ that are evenly distributed on the object plane covering an area larger than the limits of the circle. Thus, it is possible to measure how good $\hat{\mathbf{H}}(t)$ is able to represent the true homography $\mathbf{H}$ beyond the space of the control points.

Each simulation for a given camera pose is then performed in the following way: 1) An initial random n-point set $\{\mathbf{X}_i^P\}(t_{start})$ is defined inside the circular plane 2) For each iteration step t an improved set of control points $\{\mathbf{X}_i^P\}(t)$ is obtained by (11) and projected to image coordinates $\{\mathbf{x}_i\}(t)$ using the true camera pose $\mathbf{R},\mathbf{T}$ and calibration matrix $\mathbf{K}$. Then, the correspondences $\{\mathbf{x}_i(t), \mathbf{X}_i^P(t)\}$ are used to calculate $\mathbf{A}(t)$ and $c(\mathbf{A}(t))$. 3) For each t a statistically meaningful measure of the homography estimation robustness against noise is desired. Thus, 1000 runs of the homography estimation using the normalized DLT algorithm were performed[8]. In each of these runs Gaussian noise with standard deviation σ_G was added to the image coordinates for the simulation of real camera measurements $\{\tilde{\mathbf{x}}_i\}(t)$. Finally, the error $HE\left(\hat{\mathbf{H}}(t)\right)$ is calculated in each run and the average $\mu\left(HE\left(\hat{\mathbf{H}}(t)\right)\right)$ and standard deviation $\sigma\left(HE\left(\hat{\mathbf{H}}(t)\right)\right)$ of this error for all runs is computed.

As illustration of the gradient minimization process an example case of a simulation in a fronto-parallel camera pose for a 4-point configuration is presented. A Gaussian noise of $\sigma_G = 4$ pixel is added to image coordinates for the homography estimation runs. In Fig. 3 the initial object and image point configurations are shown.

The evolution of $c(\mathbf{A}(t))$ as well as $\mu\left(HE\left(\hat{\mathbf{H}}(t)\right)\right)$ and $\sigma\left(HE\left(\hat{\mathbf{H}}(t)\right)\right)$ is presented in Fig. 4. The condition number decreases drastically in the first iterations of the gradient descent, and by doing so the mean and standard deviation of $HE\left(\hat{\mathbf{H}}(t)\right)$ is also reduced. With more iterations both metrics slowly and smoothly converge to a stable minimum value.

This first result in itself is highly representative as it proves that some point configurations increase the accuracy of homography estimation methods as well as the robustness to noise and it is also possible to obtain optimized point configurations (which are better than random ones).

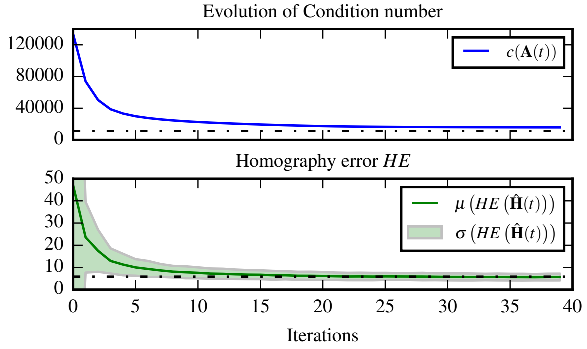

Fig. 4. (Fronto-Parallel). Evolution of the condition number $c(\mathbf{A}(t))$ as well as mean μ and standard deviation σ of the homography reprojection error $HE\left(\hat{\mathbf{H}}(t)\right)$ during gradient descent. For comparison, the dashed-dotted black line represents the mean value for an ideal 4-point square.

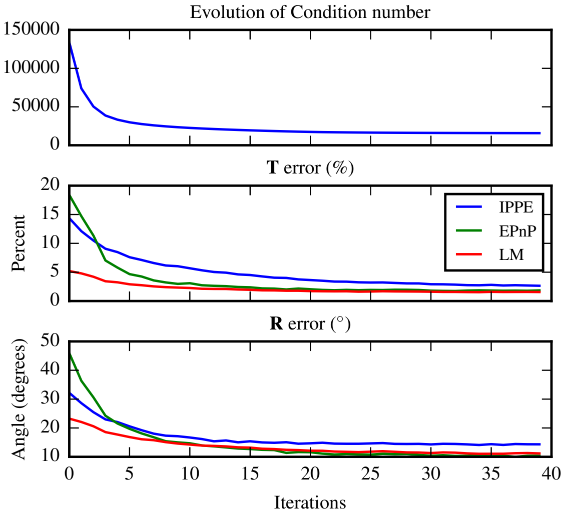

Fig. 5. (Fronto-Parallel). Comparison of the evolution of the mean errors for different PnP estimation methods during the iterative optimization process. The initial points were the same for all runs.

Motivated by the homography results, it was of interest to test if the optimization of control point configurations could improve as well the accuracy of pose estimation algorithms. Thus, three different pose estimation algorithms[9] were run at each iteration t of the optimization process, namely: 1) a non-iterative PnP method **EPnP** [8], 2) a planar pose estimation method **IPPE** [17], and 3) an iterative one based on the Levenberg-Marquardt optimization denoted as **LM**.

As in similar works [8], [17], we denote $\left(\hat{\mathbf{R}}(t), \hat{\mathbf{T}}(t)\right)$ as the estimated rotation and translation for a given camera pose at iteration t and by $(\mathbf{R},\mathbf{T})$ the true rotation and translation. The error metrics for pose estimation are defined as follows:

- $\mathrm{RE}\left(\hat{\mathbf{R}}(t)\right)$ is the rotational error (in degrees) defined as the minimal rotation needed to align $\hat{\mathbf{R}}(t)$ to $\mathbf{R}$. It is obtained from the axis-angle representation of $\hat{\mathbf{R}}(t)^T\mathbf{R}$.
- $\mathrm{TE}\left(\hat{\mathbf{T}}(t)\right) = \|\hat{\mathbf{T}}(t) - \mathbf{T}\|_2 / \|\mathbf{T}\|_2 \times 100\%$ is the relative error in translation.

[8] The homography estimation method presented in [19] and the gradient based one of OpenCV were also tested. The results almost do not differ for low point configurations to the DLT, so it was the chosen one for the experiments.

[9] For the EPnP and LM methods, the OpenCV implementations were used, and for IPPE the Python implementation provided in the author's github repository.

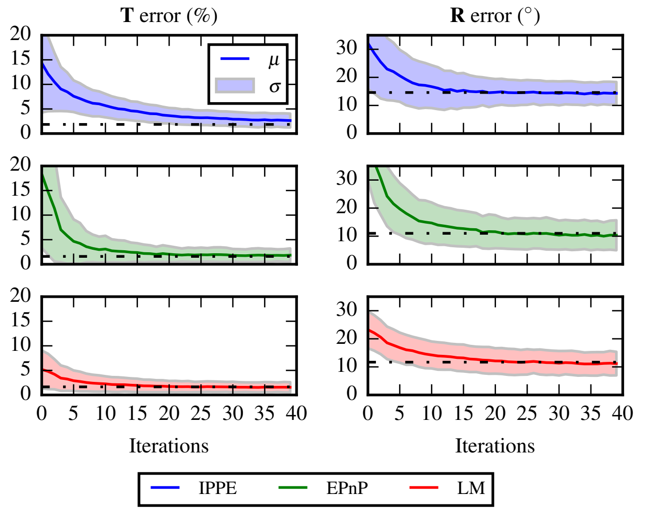

Fig. 6. (Fronto-Parallel). Mean values (colored lines) and standard deviations (filled colored areas) of translational **T** and rotational **R** error for each method. The dashed-dotted black line represents the mean value for an ideal 4-point square.

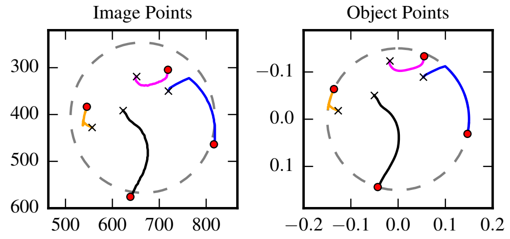

Fig. 7. (Real). Movement of control points in image and object coordinates during gradient descent for the experiment with a real camera. See our video for further details.

Similar to the homography simulation, for each iteration t, 1000 runs of the pose estimation with noisy correspondences for each of the PnP methods were performed. Then, the mean and standard deviation of RE and TE for the 1000 runs were calculated for each iteration. The PnP simulation results for the fronto-parallel case are presented in Fig. 5 comparing the performance of all methods together and in Fig. 6 details about the standard deviation of each method are shown.

A real experiment was also implemented in order to test if the simulation assumptions (Gaussian image noise and perfect intrinsics) may affect the results in practical applications[10]. A computer screen was used as the planar fiducial marker to dynamically display the points during gradient descent. A set of 4 circles was displayed for each iteration of the optimization. These circles were then captured by a PointGrey Blackfly camera[11] and detected using a Hough transform based circle detector. We performed 100 detections for each gradient descent iteration. An Optitrack system was used to measure the ground truth pose of the camera relative to the marker screen. The results of running the optimization process for a set of 4 random initial points are shown in figures 7, 8 and 9.

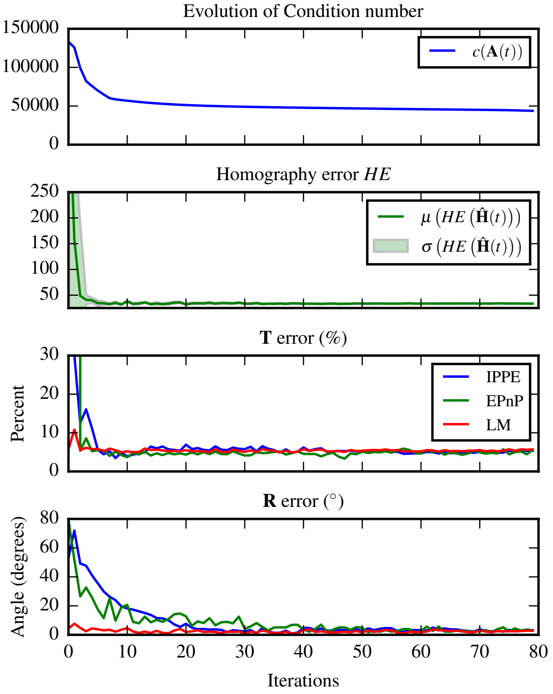

Fig. 8. (Real). Evolution of the condition number and the homography reprojection error during gradient descent using a real camera.

Next, the relationship between badly configured points and optimized points was studied. For each camera pose in the distribution of Fig. 2, 100 different initial random n-point configurations with $n \in \{4,5,6,7,8\}$ were simulated and the optimization process was performed. In this case, only the initial and final values of the point configuration metrics are stored. Thus, it is possible to compare the methods based on *ill-conditioned* (random initial points) and *well-conditioned* (after optimization) point configurations. In Fig. 10 the results for the homography estimation are presented and in Fig. 11 the results of the pose estimation. Finally, in figures 12 and 13, the final point configurations for all the camera poses are shown as a 2D histogram and some example configurations are shown for the 4-point and 5-point case.

B. Discussion

In the results, it is observed that control point configurations have a strong effect on the accuracy of homography and planar PnP methods. There are indeed optimized configurations which are better than random and it is possible to find them using out method.

For the 4-point case, our empirical results show that a square-like shape is the most common minima and a very stable and robust configuration for all camera poses (see Fig. 12) and as shown on Fig. 13 even for the 5-point case the corners of a square-like shape are common. The optimized point configurations do not show any strong dependency with the pose of the camera (besides scale and image limits), it is mainly related to the distribution of the points in camera

[10] A video of this experiment: https://youtu.be/a6lDrwgqNmY.

[11] Camera parameters: size $1288 \times 964\,[pixel^2]$, intrinsic parameters $\mathbf{K} = [1070.82, 0, 647.98; 0, 1071.20, 488.27; 0, 0, 1]$.

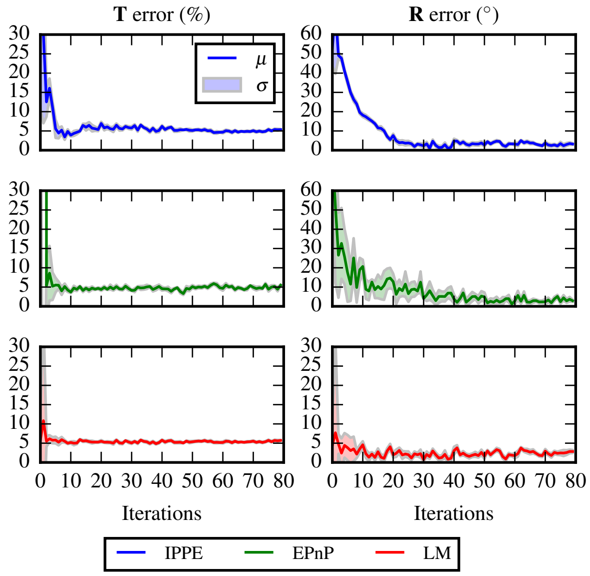

Fig. 9. (Real). Detailed view of the standard deviation of each method represented by the filled, lightly colored areas.

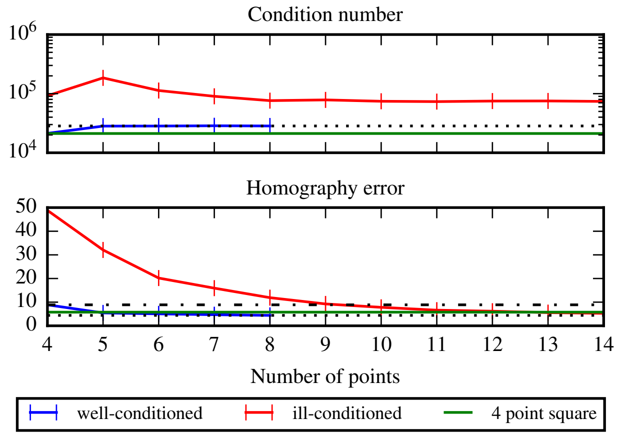

Fig. 10. Robustness (cond. num.) and accuracy (homography error) dependent on the number of points for well- and ill-conditioned point configurations as well as an ideal 4-point square (green line).

image coordinates since they are driven to distribute in space and they tend to increase the distance to each other.

On the first iterations of the optimization is when the increase in accuracy is stronger, which means that the condition number is a good optimization objective. For example, the improvement in accuracy from a square-like configuration to a perfect square is very small, but the increase of accuracy from random points to the square-like shapes obtained on the first iterations of the optimization is radical.

The smaller the number of control points the more is the relative improvement on the estimates for all of the evaluated methods. For example, the accuracy using 4 points is always better than random point configurations with more points $4 < n \leq 9$ as can be seen in Fig. 10 for homography and Fig. 11 for PnP. Thus, the configuration of the control points has more effect on the accuracy than the number of control points.

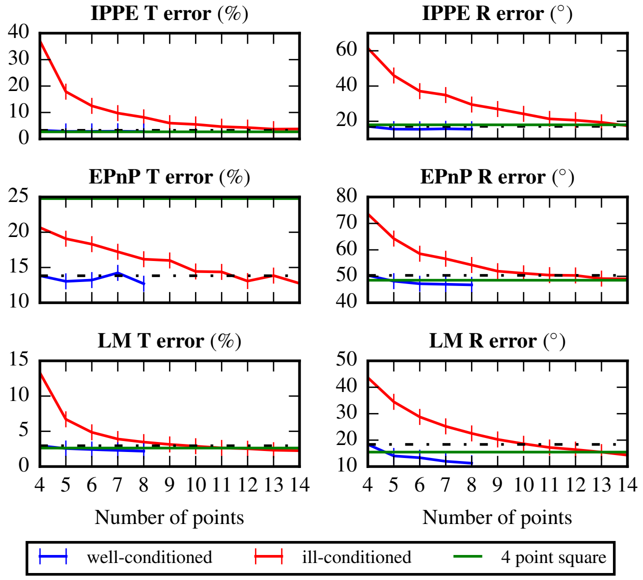

Fig. 11. Comparison of different pose stimation methods for different numbers of control points for well- and ill-conditioned point configurations as well as an ideal 4-point square (green line).

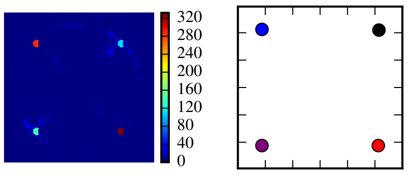

Fig. 12. Left: 2D histogram of final 4-point configurations for all camera poses. Right: One representative final point configuration. In object coordinates.

The improvement in the EPnP and IPPE methods is more pronounced than for LM, which is in itself an interesting result since those methods take considerable less computation time. For well-configured points, the methods converge to similar error values (see Fig. 5, 6 and Fig. 8) and both mean and variance are reduced, this means that well-conditioned points can be used for fair comparison of pose estimation algorithms. LM also has increased accuracy although our optimization objective is not directly related to the minimization of the reprojection error, this shows the importance of having a good initial guess. The results of the real experiment closely match the simulations.

VI. Conclusions and Future Work

A method for obtaining optimized control points for homography estimation is presented. The lower the number of control points the more the point configuration has an influence on the accuracy of homography and PnP estimation methods. Our empirical results show that a square is a very stable and robust configuration for all camera poses. Optimized points configurations follow simple rules, they are driven to distribute in space and they tend to increase the distance to each other, this includes the optimized 4 point configuration as a subset. Finally, we found that there

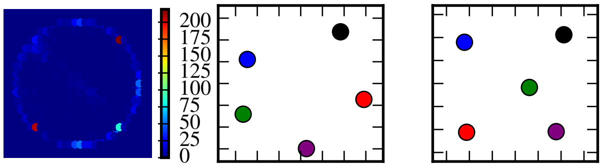

Fig. 13. Left: 2D histogram of final 5-point configurations for all camera poses. Middle/right: Two representative final point configurations. In object coordinates.

is almost no difference in accuracy between IPPE and LM when optimized point configurations are used. In future work, we will try to generalize the results to non-planar point configurations and use other optimization metrics such as the trace of the posterior covariance matrix in the reprojection error which is commonly used in the optimal sensor placement research field.

References

[1] R. Hartley, "In defence of the 8-point algorithm," in *Proc. of IEEE Int. Conf. on Computer Vision*, 1997.

[2] P. Chen and D. Suter, "Error analysis in homography estimation by first order approximation tools: A general technique," *Journal of Mathematical Imaging and Vision*, 2009.

[3] E. Marchand, H. Uchiyama, and F. Spindler, "Pose Estimation for Augmented Reality: A Hands-On Survey," *IEEE Trans. on Vis. and Comput. Graphics*, 2016.

[4] D. Oberkampf, D. F. DeMenthon, and L. S. Davis, "Iterative Pose Estimation Using Coplanar Feature Points," *Computer Vision and Image Understanding*, 1996.

[5] C. P. Lu, G. D. Hager, and E. Mjolsness, "Fast and globally convergent pose estimation from video images," *IEEE Trans. Pattern Anal. Mach. Intell.*, 2000.

[6] V. Garro, F. Crosilla, and A. Fusiello, "Solving the PnP problem with anisotropic orthogonal procrustes analysis," *3DIMPVT*, 2012.

[7] G. Schweighofer and A. Pinz, "Globally Optimal O(n) Solution to the PnP Problem for General Camera Models," *BMVC*, 2008.

[8] V. Lepetit, F. Moreno-Noguer, and P. Fua, "EPnP: An accurate O(n) solution to the PnP problem," *International Journal of Computer Vision*, 2008.

[9] J. A. Hesch and S. I. Roumeliotis, "A direct least-squares (DLS) method for PnP," in *IEEE Int. Conf. on Computer Vision*, 2011.

[10] S. Li, C. Xu, and M. Xie, "A robust O(n) solution to the perspective-n-point problem," *IEEE Trans. Pattern Anal. Mach. Intell.*, 2012.

[11] Y. Zheng, Y. Kuang, S. Sugimoto, K. Astrom, and M. Okutomi, "Revisiting the PnP problem: A fast, general and optimal solution," in *Proc. of the IEEE Int. Conf. on Computer Vision*, 2013.

[12] L. Kneip, H. Li, and Y. Seo, "UPnP: An Optimal O(n) Solution to the Absolute Pose Problem with Universal Applicability," in *European Conference on Computer Vision*, 2014.

[13] G. Nakano, "Globally Optimal DLS Method for PnP Problem with Cayley parameterization," in *BMVC*, 2015.

[14] P. Sturm, "Algorithms for Plane-Based Pose Estimation," *IEEE Conf. on Computer Vision and Pattern Recognition*, 2000.

[15] Z. Zhang, "A Flexible New Technique for Camera Calibration," *IEEE Trans. Pattern Anal. Mach. Intell.*, 2000.

[16] G. Schweighofer and A. Pinz, "Robust pose estimation from a planar target," *IEEE Trans. Pattern Anal. Mach. Intell.*, 2006.

[17] T. Collins and A. Bartoli, "Infinitesimal plane-based pose estimation," *International Journal of Computer Vision*, 2014.

[18] A. Z. Richard Hartley, *Multiple View Geometry*, 2nd ed. Cambridge University Press, 2004.

[19] M. J. Harker and P. L. O'Leary, "Computation of Homographies," in *BMVC*, 2005.

[20] P. Rangarajan and P. Papamichalis, "Estimating homographies without normalization," in *Proc. of Int. Conf. on Image Processing*, 2009.

[21] V. Willert, "Optical indoor positioning using a camera phone," in *Proc. of the 2010 int. conf. on indoor positioning and indoor navigation*, 2010.

[22] D. H. S. Chung, M. L. Parry, P. A. Legg, I. W. Griffiths, R. S. Laramee, and M. Chen, "Visualizing multiple error-sensitivity fields for single camera positioning," *Computing and Visualization in Science*, 2014.

[23] G. W. Stewart, "Perturbation theory for the singular value decomposition," Tech. Rep., 1998.

[24] G. H. Golub and C. F. Van Loan, *Matrix Computations*, 4th ed., 2013.

[25] D. Maclaurin, "Modeling, inference and optimization with composable differentiable procedures," Tech. Rep., 2016.

[26] L. B. Rall, "Automatic differentiation: Techniques and applications," 1981.

[27] T. Tollenaere, "Supersab: fast adaptive back propagation with good scaling properties," pp. Neural Networks 3(5), 561–573, 1990.