Camera Pose Estimation with Semantic 3D Model

Vincent Gaudillière, Gilles Simon, Marie-Odile Berger

To cite this version:

Vincent Gaudillière, Gilles Simon, Marie-Odile Berger. Camera Pose Estimation with Semantic 3D Model. IROS 2019 - 2019 IEEE/RSJ International Conference on Intelligent Robots and Systems, Nov 2019, Macau, Macau SAR China. 10.1109/IROS40897.2019.8968180 . hal-02181962

HAL Id: hal-02181962

https://hal.science/hal-02181962

Submitted on 12 Jul 2019

Camera Pose Estimation with Semantic 3D Model

Vincent Gaudillière[1], Gilles Simon[1], Marie-Odile Berger[1]

Abstract— **In computer vision, estimating camera pose from correspondences between 3D geometric entities and their projections into the image is a widely investigated problem. Although most state-of-the-art methods exploit simple primitives such as points or lines, and thus require dense scene models, the emergence of very effective CNN-based object detectors in the recent years have paved the way to the use of much lighter 3D models composed solely of a few semantically relevant features. In that context, we propose a novel model-based camera pose estimation method in which the scene is modeled by a set of virtual ellipsoids. We show that 6-DoF camera pose can be determined by optimizing only the three orientation parameters, and that at least two correspondences between 3D ellipsoids and their 2D projections are necessary in practice. We validate the approach on both simulated and real environments.**

I. INTRODUCTION

Camera pose estimation is a fundamental task in computer vision. In this problem, it is necessary to build and maintain a representation of the environment in which the observer (*e.g.* a robot) operates. In practice, this knowledge is most often presented in the form of a 3D model, with respect to which the camera is positioned [1]. When the scene is modeled by a point cloud, camera pose can be estimated given at least three correspondences between a 3D point and its projection into the image (P3P problem) [2]. To achieve higher accuracy, most methods consider an arbitrary number $n > 3$ of 2D-3D correspondences (PnP) [3], [4]. There are also methods for models made up of lines (PnL), or a mixture of points and lines [5]. These models are usually built from either Structure from Motion or SLAM techniques [6].

In the recent years, significant progress have been made in automatic object detection thanks to methods based on convolutional neural networks such as R-CNN [7], [8], [9], SSD [10], or YOLO [11], [12], [13]. This qualitative leap has led to the emergence of new approaches to solve traditional computer vision problems. In particular, extraction and matching of high-level features (objects), instead of the traditional low-level primitives (visual keypoints, line segments), is already at the basis of several methods in the literature. Modeling object projections by virtual ellipses allowed Crocco *et al.* to propose a closed-form solution for SfM reconstruction of the scene in the form of an ellipsoid cloud [14]. However, this method is limited to the case of an orthographic projection, as well as its extension integrating CAD object models for higher reconstruction accuracy [15]. Perspective projection is taken into account in [16], but this SLAM method proposes to minimize a geometric reprojection error as a function of the six degrees of freedom of the camera, based on an initial solution provided by odometry measurements. Other SLAM techniques take benefit of object landmarks present in the scene to achieve higher accuracy and robustness in comparison with semantically-unaware methods, yet still requiring visual [17] or sensor-based [18] 6-DoF estimates of camera poses.

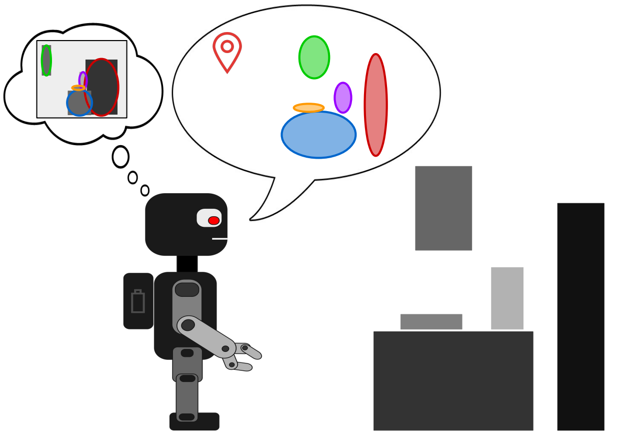


Fig. 1. Our system is able to estimate its position from objects detected in the image. Object detections are modeled by ellipses, considered as projections of a known 3D scene model composed of virtual ellipsoids.

In this paper, we propose a model-based method for camera pose estimation from ellipse-ellipsoid correspondences. Our algorithm assumes that the scene is modeled by a set of virtual ellipsoids corresponding to objects of interest. We use the approach described by Rubino *et al.* in [19] to build such a model from at least three calibrated perspective cameras. Object projections, once detected in the image then fitted by ellipses, are used to estimate the camera pose. Wokes and Palmer proposed a method for calculating the pose of an object modelled by a spheroid (ellipsoid with two equal semi-diameters) [20], [21]. The authors showed that, considering perspective projection, the spheroid pose estimation problem admits only two distinct solutions. In the more general case of ellipsoid, an equation of the same problem was proposed by Eberly [22], without however an explicit method for calculating solutions.

In this article, we take up the formalism introduced by Eberly, and show that the problem of ellipsoid pose estimation from its projected ellipse is equivalent to a second one, in which only the orientation of the ellipsoid is involved (Section II). The ellipsoid position can then be uniquely in-

*This work was supported by the DGA/DGE RAPID EVORA project.

[1] Université de Lorraine, CNRS, Inria, LORIA, F-54000 Nancy, France
{firstname.lastname}@inria.fr

ferred from its orientation (Section III-A). These theoretical considerations allow us to propose a method for estimating 6-DoF camera pose from correspondences between 2D ellipses and 3D ellipsoids, in the form of an optimization problem whose only parameters are the camera orientation ones. We numerically highlight the fact that the pose cannot be uniquely determined from a single correspondence (Section III-B), before formulating the problem for an arbitrary number of ellipsoids (Section III-C). Finally, robustness of the method is evaluated, and compared to the state of the art (Section IV).

II. PROBLEM REDUCTION

First of all, we focus on the problem of camera pose estimation with a scene composed of a single ellipsoid. In fact, we consider the equivalent problem that consists in calculating the ellipsoid pose in the camera frame, and show that it can be reduced to only orientation determination. In line with the work of Wokes and Palmer about spheroids [21], we propose here an analysis of the general case (arbitrary ellipsoid). However, the formalism that we use is different from the one of the reference, since the latter is based on the assumption that the ellipsoid has an axis of symmetry.

A. The Cone Alignment Equation

Unless otherwise stated, all the variables introduced below are expressed in the camera coordinate frame.

Following the notations introduced in [22] and presented in Fig. 2, we consider an ellipsoid defined by

$$(\mathbf{X}-\mathbf{C})^\top A(\mathbf{X}-\mathbf{C}) = 1$$

where $\mathbf{C}$ is the center of the ellipsoid, A is a real positive definite matrix characterizing its orientation and size, and $\mathbf{X}$ is any point on it.

Given a center of projection $\mathbf{E}$ and a projection plane of normal $\mathbf{N}$ which does not contain $\mathbf{E}$, the projection of the ellipsoid is an ellipse of center $\mathbf{K}$ and of semi-diameters a et b. Ellipse's principal directions are represented by unit-length vectors $\mathbf{U}$ and $\mathbf{V}$, such that $\{\mathbf{U},\mathbf{V},\mathbf{N}\}$ is an orthonormal set.

1) Projection Cone: The "projection cone" refers to the cone of vertex $\mathbf{E}$ tangent to the ellipsoid. According to [22], it is defined by the matrix

$$B \overset{def}{=} A\boldsymbol{\Delta}\boldsymbol{\Delta}^\top A - (\boldsymbol{\Delta}^\top A\boldsymbol{\Delta} - 1)A$$

where $\boldsymbol{\Delta} = \mathbf{E} - \mathbf{C}$, so that the points $\mathbf{X}$ on the projection cone are those who satisfy the equation $(\mathbf{X}-\mathbf{E})^\top B(\mathbf{X}-\mathbf{E}) = 0$. Note that B is a real, symmetric and invertible matrix which has two eigenvalues of the same sign and the third one of the opposite sign.

2) Backprojection Cone: The "backprojection cone" refers to the cone generated by the lines passing through $\mathbf{E}$ and any point on the ellipse. Eberly shows that such a cone is characterized by the matrix B' defined as follows

$$B' \overset{def}{=} P^\top MP - Q$$

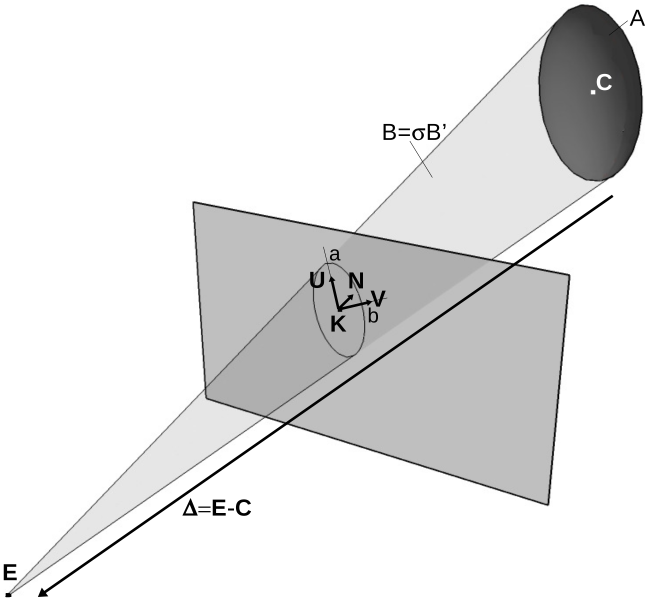


Fig. 2. Illustrating the projection plane, projection center, ellipsoid and projected ellipse.

where

$$M \overset{def}{=} \mathbf{U}\mathbf{U}^\top/a^2 + \mathbf{V}\mathbf{V}^\top/b^2$$
$$\mathbf{W} \overset{def}{=} \mathbf{N}/(\mathbf{N}\cdot(\mathbf{K}-\mathbf{E}))$$
$$P \overset{def}{=} I - (\mathbf{K}-\mathbf{E})\mathbf{W}^\top$$
$$Q \overset{def}{=} \mathbf{W}\mathbf{W}^\top$$

Here again, the points $\mathbf{X}$ on the backprojection cone are those who meet $(\mathbf{X}-\mathbf{E})^\top B'(\mathbf{X}-\mathbf{E}) = 0$, and B' has the same properties as B (real, symmetric, invertible with signature (2,1) or (1,2)).

3) Alignment Equation: Given an ellipsoid, a central projection (center and plane), and an ellipse on the projection plane, the ellipse is the projection of the ellipsoid if and only if the projection and backprojection cones are aligned [22], *i.e.* if and only if there is a non-zero scalar σ such that $B = \sigma B'$:

$$A\boldsymbol{\Delta}\boldsymbol{\Delta}^\top A - (\boldsymbol{\Delta}^\top A\boldsymbol{\Delta} - 1)A = \sigma B' \tag{1}$$

Moreover, one can define a coordinate frame in which the quadratic form related to the ellipsoid is represented by a diagonal matrix D (with strictly positive diagonal entries). By noting R the rotation between that coordinate frame and the one of the image, we have $A = R^\top DR$.

Thus, solving the ellipsoid pose estimation problem consists in solving the following equation, of unknowns $(R, \boldsymbol{\Delta}, \sigma)$:

$$DR\boldsymbol{\Delta}\boldsymbol{\Delta}^\top R^\top D - (\boldsymbol{\Delta}^\top R^\top DR\boldsymbol{\Delta} - 1)D = \sigma RB'R^\top$$

In that formulation, R characterizes the orientation of the ellipsoid, $\boldsymbol{\Delta}$ characterizes its position, and σ is an additional scalar parameter.

B. Equivalence Theorem

The ellipsoid pose estimation problem therefore involves three variables. In this section, we show that $\mathbf{\Delta}$ and σ are actually secondary variables. For this, we show that the problem formalized in section II-A is equivalent to a second problem, in which the only unknown is R.

In what follows, saying that an ellipsoid A is a generator of the cone B' will refer to the situation where there is a vector $\mathbf{\Delta}$ and a non-zero scalar σ such that the triplet $(A, \mathbf{\Delta}, \sigma)$ is solution of (1).

Theorem 1. *A is a generator of B' if and only if the discriminant of the generalized characteristic polynomial of the pair $\{A, B'\}$ is zero.*

Proof. • **Let's suppose that there is a vector $\mathbf{\Delta}$ and a non-zero scalar σ so that the triplet $(A, \mathbf{\Delta}, \sigma)$ is solution of (1).** Then multiplying (1) on the right by $\mathbf{\Delta}$ (see Appendix 1) leads to

$$A\mathbf{\Delta} = \sigma B'\mathbf{\Delta} \tag{2}$$

From there, finding the pairs $(\sigma, \mathbf{\Delta})$ that satisfy (2) amounts to solving a generalized eigenvalue problem [23]. Such σ values are called *generalized eigenvalues* of the couple $\{A, B'\}$, and their corresponding $\mathbf{\Delta}$ are called their *generalized eigenvectors*. In particular, the generalized eigenvalues of $\{A, B'\}$ are the roots of the *generalized characteristic polynomial* $P_{\{A,B'\}}(x) = det(A - xB')$. Yet since B' is invertible, we can easily notice that the generalized eigen elements of the pair $\{A, B'\}$ are the same as the eigen elements of the matrix $B'^{-1}A$. We can then observe that $Q(x) = \mu x^2 - (\mu+1)\sigma x + \sigma^2$, where $\mu = 1 - \mathbf{\Delta}^\top A\mathbf{\Delta}$, is an annihilator polynomial of $B'^{-1}A$ (see Appendix 2). Since Q is of degree 2, we can infer that $B'^{-1}A$, and thus $\{A, B'\}$, have at most two distinct eigenvalues. In other words, the polynomial $P_{\{A,B'\}}$ has at most two distinct roots, thus has its discriminant equal to zero (see Appendix 3).

• **Now let's suppose that the discriminant of $P_{\{A,B'\}}$ is zero**. We can first notice that, since A is positive definite and B' is symmetric, the couple $\{A, B'\}$ has the following properties [23]:

1) the generalized eigenvalues are real,
2) the reducing subspaces are of the same dimension as the multiplicity of the associated eigenvalues,
3) the generalized eigenvectors form a basis of $\mathbb{R}^3$, and those with distinct eigenvalues are A-orthogonal.

Since the discriminant of $P_{\{A,B'\}}$ is equal to zero, the couple $\{A, B'\}$ has at most two distinct eigenvalues. Moreover, if the couple had only one eigenvalue of multiplicity 3 called σ_0, then according to property 2) above, we would have $dim(Ker(A - \sigma_0 B')) = 3$, *i.e.* $A = \sigma_0 B'$, which is impossible because A represents an ellipsoid while B' represents a cone. So **the couple has exactly two distinct generalized eigenvalues.**

Let's then denote σ_1 (multiplicity 1) and σ_2 (multiplicity 2) these two eigenvalues. Observing that $\frac{1}{\sigma_1}$ and $\frac{1}{\sigma_2}$ are the generalized eigenvalues of the couple $\{B', A\}$, we can write, according to [24] (Theorem 3)

$$\forall \mathbf{X} \in \mathbb{R}^3 \backslash \{\mathbf{0}\},\ min(\frac{1}{\sigma_1}, \frac{1}{\sigma_2}) \leq \frac{\mathbf{X}^\top B'\mathbf{X}}{\mathbf{X}^\top A\mathbf{X}} \leq max(\frac{1}{\sigma_1}, \frac{1}{\sigma_2})$$

If σ_1 and σ_2 were of the same sign, then $\forall \mathbf{X} \in \mathbb{R}^3 \backslash \{\mathbf{0}\}$, $\mathbf{X}^\top B'\mathbf{X}$ would be of that sign (since $\mathbf{X}^\top A\mathbf{X} > 0$). Yet, it is impossible since B' is neither positive nor negative definite (cone). We thus conclude that **the two distinct eigenvalues are of opposite signs.**

Let's denote $\mathbf{\Delta}, \mathbf{U}, \mathbf{V}$ three eigenvectors associated with σ_1 and σ_2. Therefore, by definition

$$\begin{aligned} A\mathbf{\Delta} &= \sigma_1 B'\mathbf{\Delta} \\ A\mathbf{U} &= \sigma_2 B'\mathbf{U} \\ A\mathbf{V} &= \sigma_2 B'\mathbf{V} \end{aligned}$$

Denoting $B_A(\mathbf{X})$ the matrix $A\mathbf{X}\mathbf{X}^\top A - (\mathbf{X}^\top A\mathbf{X} - 1)A$, we can observe that $\forall \mathbf{X} \in \mathbb{R}^3$

$$\begin{aligned} B_A(\mathbf{X})\mathbf{X} &= A\mathbf{X}(\mathbf{X}^\top A\mathbf{X}) - \mathbf{X}^\top A\mathbf{X}A\mathbf{X} + A\mathbf{X} \\ &= (\mathbf{X}^\top A\mathbf{X})A\mathbf{X} - \mathbf{X}^\top A\mathbf{X}A\mathbf{X} + A\mathbf{X} \\ &= A\mathbf{X} \end{aligned}$$

In particular, $B_A(\mathbf{\Delta})\mathbf{\Delta} = A\mathbf{\Delta}$, thus

$$B_A(\mathbf{\Delta})\mathbf{\Delta} = \sigma_1 B'\mathbf{\Delta}$$

Then, right-multiplying $B_A(\mathbf{\Delta})$ by U, after noticing that $\mathbf{\Delta}$ and $\mathbf{U}$ are A-orthogonal (see property 3) above), leads to

$$\begin{aligned} B_A(\mathbf{\Delta})\mathbf{U} &= A\mathbf{\Delta}\mathbf{\Delta}^\top A\mathbf{U} - (\mathbf{\Delta}^\top A\mathbf{\Delta} - 1)A\mathbf{U} \\ &= A\mathbf{\Delta}(\mathbf{\Delta}^\top A\mathbf{U}) + (1 - \mathbf{\Delta}^\top A\mathbf{\Delta})A\mathbf{U} \\ &= (1 - \mathbf{\Delta}^\top A\mathbf{\Delta})A\mathbf{U} \\ &= (1 - \mathbf{\Delta}^\top A\mathbf{\Delta})\sigma_2 B'\mathbf{U} \end{aligned}$$

From there, we can choose the norm of $\mathbf{\Delta}$ such that $(1 - \mathbf{\Delta}^\top A\mathbf{\Delta})\sigma_2 = \sigma_1$. Indeed, all we have to do is choosing $\|\mathbf{\Delta}\|$ such that $\mathbf{\Delta}^\top A\mathbf{\Delta} = 1 - \frac{\sigma_1}{\sigma_2}$, which is enabled by the fact that σ_1 and σ_2 are of opposite signs, thus $1 - \frac{\sigma_1}{\sigma_2} > 1 > 0$. Choosing such a $\mathbf{\Delta}$, we obtain

$$B_A(\mathbf{\Delta})\mathbf{U} = \sigma_1 B'\mathbf{U}$$

and, in the same way

$$B_A(\mathbf{\Delta})\mathbf{V} = \sigma_1 B'\mathbf{V}$$

Since $(\mathbf{\Delta}, \mathbf{U}, \mathbf{V})$ form a basis of $\mathbb{R}^3$, we finally have

$$B_A(\mathbf{\Delta}) \overset{def}{=} A\mathbf{\Delta}\mathbf{\Delta}^\top A - (\mathbf{\Delta}^\top A\mathbf{\Delta} - 1)A = \sigma_1 B'$$

So, A is part of a triplet of solution $(A, \mathbf{\Delta}, \sigma_1)$ of (1). □

To conclude, ellipsoid pose estimation can be reduced to only orientation estimation. Orientations R that are solutions are those that annihilate the discriminant of the generalized characteristic polynomial of the couple $\{A = R^\top DR, B'\}$.

III. Camera Pose Estimation Method

In this section, we show that the position can then be uniquely determined from the orientation (III-A), then we numerically highlight the fact that there is a continuum of solutions for R (III-B). We therefore formulate the problem of camera pose estimation with a scene composed of at least two virtual ellipsoids (III-C).

A. From Orientation to Position

Theorem 1 has shown that $\mathbf{\Delta}$ and σ are secondary variables of (1). In this section, we also show that they can be uniquely determined from A and B'.

Corollary 1. *If A is a generator of B', then (i) the couple {A,B'} has exactly two distinct eigenvalues of opposite signs, (ii) σ is the eigenvalue of multiplicity 1, and (iii) $\mathbf{\Delta}$ is an eigenvector associated with σ and is unique.*

Proof. We have seen that (2) has two solutions. Let's denote them $(\sigma_1, \mathbf{\Delta}_1)$ and $(\sigma_2, \mathbf{\Delta}_2)$, such that σ_i is the eigenvalue of multiplicity i and $\|\mathbf{\Delta}_i\| = 1$. Let's suppose now that there is $k \in \mathbb{R}^*$ such that $(A, \sigma_2, k\mathbf{\Delta}_2)$ is solution of (1). We therefore have

$$A - \sigma_2 B' = MA$$

denoting $M = k^2(\mathbf{\Delta}_2^\top A \mathbf{\Delta}_2 I - A\mathbf{\Delta}_2\mathbf{\Delta}_2^\top)$, where I is the identity matrix. According to the property 2) of the proof of Theorem 1, $dim(Ker(A - \sigma_2 B')) = 2$, thus, since A is invertible, $dim(Ker(M)) = 2$. However, we observe that

$$\begin{aligned} \forall \mathbf{X} \perp \mathbf{\Delta}_2,\ M\mathbf{X} &= k^2\mathbf{\Delta}_2^\top A\mathbf{\Delta}_2\mathbf{X} - k^2 A\mathbf{\Delta}_2\mathbf{\Delta}_2^\top\mathbf{X} \\ &= k^2\mathbf{\Delta}_2^\top A\mathbf{\Delta}_2\mathbf{X} - k^2 A\mathbf{\Delta}_2(\mathbf{\Delta}_2 \cdot \mathbf{X}) \\ &= k^2\mathbf{\Delta}_2^\top A\mathbf{\Delta}_2\mathbf{X} \end{aligned}$$

So, since A is positive definite, $M\mathbf{X} \neq 0$ when $\mathbf{X} \neq 0$. Therefore, defining $\mathbf{\Delta}_2^\perp = \{\mathbf{X} \in \mathbb{R}^3 / \mathbf{X} \perp \mathbf{\Delta}_2\}$ the subspace of dimension 2 orthogonal to $\mathbf{\Delta}_2$, and observing that the previous inequality means $\mathbf{\Delta}_2^\perp \cap Ker(M) = \{0\}$ thus $dim(\mathbf{\Delta}_2^\perp) + dim(Ker(M)) \leq 3$, we end up with a contradiction since $dim(\mathbf{\Delta}_2^\perp) = dim(Ker(M)) = 2$.

As a result, triplets $(A, \sigma_2, k\mathbf{\Delta}_2)$ cannot be solutions of (1), thus possible solutions are necessarily written $(A, \sigma_1, k\mathbf{\Delta}_1)$, where $k \in \mathbb{R}^*$. Equation (1) is then written

$$k^2(A\mathbf{\Delta}_1\mathbf{\Delta}_1^\top A - \mathbf{\Delta}_1^\top A\mathbf{\Delta}_1 A) = \sigma_1 B' - A$$

The two matrices shown above are proportional, and k^2 is their proportionality constant. Finally, the only possible k is the one that allows the center of the ellipsoid to be in front of the camera (chirality constraint). □

Corollary 1 states that σ and $\mathbf{\Delta}$ can be calculated unambiguously from A. In practice, the discriminant of $P_{\{A,B'\}}$ is not always equal to 0. In this case, the ratios between the eigenvalues of $\{A, B'\}$ are first calculated, then the two values whom ratio is closest to 1 are determined, with σ being defined as the third value. Then, we define $\mathbf{\Delta}_0$ as the eigenvector of norm 1 associated with σ, and we set $M = \sigma B' - A$ and $N = A\mathbf{\Delta}_0\mathbf{\Delta}_0^\top A - \mathbf{\Delta}_0^\top A\mathbf{\Delta}_0 A$. We then define K_2 the matrix whose general term is $M(i,j)/N(i,j)$, then $vec(K_2)$ the vector formed by the six entries of its upper triangular part (K_2 is symmetric by construction). Finally, k^2 is calculated as its average: $k^2 = \overline{vec(K_2)}$, and the sign of k is determined by applying the chirality constraint.

B. Poses from One Ellipsoid

We define $\mathscr{R}_w = (O; \mathscr{B}_w)$ the world coordinate frame, and $\mathscr{R}_c = (E; \mathscr{B}_c)$ the camera coordinate frame, with $\mathscr{B}_w$ and $\mathscr{B}_c$ two direct orthonormal bases. The matrix representation, in the base $\mathscr{B}_w$, of the quadratic form associated with the ellipsoid is called A_w, and its representation in the base $\mathscr{B}_c$ is called A_c. Similarly, B'_c is the matrix representation, in the base $\mathscr{B}_c$, of the quadratic form associated with the backprojection cone. In addition, we denote $\mathbf{\Delta}_w$ the expression, in the base $\mathscr{B}_w$, of the vector connecting the center of the ellipsoid to the center of the camera, and $\mathbf{\Delta}_c$ its expression in the base $\mathscr{B}_c$. Finally, we refer to ${}^wR_c$ as the rotation matrix between $\mathscr{B}_w$ and $\mathscr{B}_c$, by which we can write

$$\left\{ \begin{array}{l} A_c = {}^wR_c^\top A_w {}^wR_c \\ \mathbf{\Delta}_w = {}^wR_c \mathbf{\Delta}_c \end{array} \right.$$

Assuming that A_w is known (model), the problem consists then in finding an orientation ${}^wR_c$ of the camera such that $discriminant(P_{\{A_c, B'_c\}}) = 0$. We can therefore see it as an optimization problem. It is thus necessary to assume that an initial estimate of the rotation between the world and camera coordinate frames, characterized by the Euler angles $\mathbf{\Theta}_0 = (\theta_0^{(1)}, \theta_0^{(2)}, \theta_0^{(3)})$, is known (calculated for example using vanishing points). The cost function

$$f_{discr}(\mathbf{\Theta}) = |discriminant(P_{\{A_c(\mathbf{\Theta}), B'_c\}})|^2$$

is minimized using a Levenberg-Marquardt algorithm.

An illustration on a synthetic case of the solutions to the problem is provided in Fig. 3. A set of orientations is computed starting from initial orientations obtained by discretizing the space of Euler angles with a 30 degree step in each direction. For each orientation, the position of the center of the ellipsoid in the camera frame was calculated as described in section III-A. Fig. 3 shows the locus of solutions for the problem of pose estimation from a single ellipsoid.

C. Pose from N Ellipsoids

Since there are an infinite number of solutions with only one ellipsoid, we now consider a scene model composed of $N \geq 2$ virtual ellipsoids. Following the previous notation philosophy, an exponent ${}^{(i)}$ is added to the quantities related to the ellipsoid i, so that for any i between 1 and N, we can write

$$\left\{ \begin{array}{l} A_c^{(i)} = {}^wR_c^\top A_w^{(i)} {}^wR_c \\ \mathbf{\Delta}_w^{(i)} = {}^wR_c \mathbf{\Delta}_c^{(i)} \end{array} \right. \tag{3}$$

Here again, the matrices $A_w^{(i)}$ are assumed to be known (model), and the camera orientation ${}^wR_c$ is characterized by

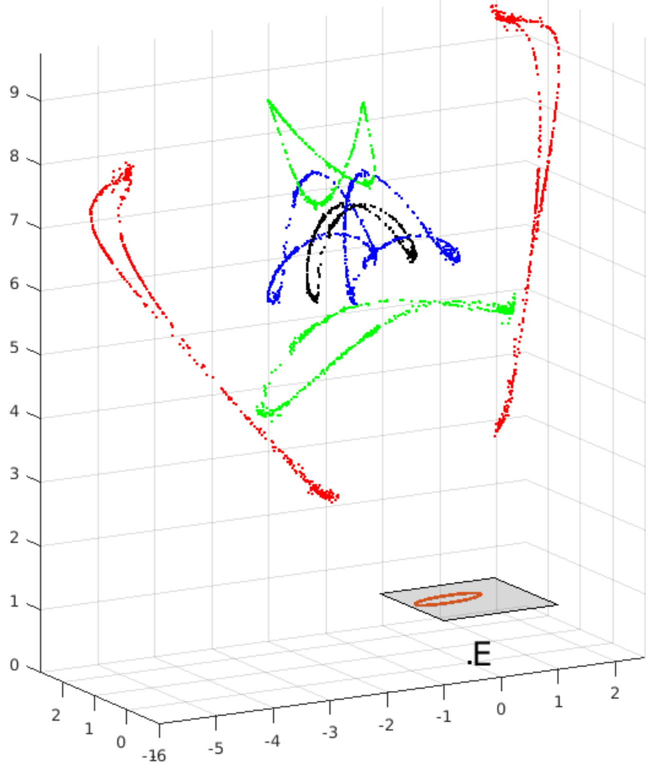


Fig. 3. Illustration in the camera frame of the solutions to the problem of pose estimation with a single ellipsoid, obtained after minimizing the f_{discr} function from 1728 different initial orientations. In black: centers of the reconstructed ellipsoids. In red, green, and blue: endpoints of their principal axes. In orange: projected ellipse in the image plane. E: camera center.

Euler angles $\mathbf{\Theta} = (\theta^{(1)}, \theta^{(2)}, \theta^{(3)})$. The new cost function to be minimized is then defined as follows:

$$f'_{discr}(\mathbf{\Theta}) = \sum_{i=1}^{N} |discriminant(P_{\{A_c^{(i)}(\mathbf{\Theta}),B_c'^{(i)}\}})|^2$$

a) Orientation computation: Starting from an initial solution $\mathbf{\Theta}_0$, we apply a Levenberg-Marquardt algorithm to minimize the function $f'_{discr}(\mathbf{\Theta})$. The obtained parameters $\mathbf{\Theta}_f$ allow to define a new camera orientation ${}^{w}R_c(\mathbf{\Theta}_f)$.

b) Position computation: We call $C_w^{(i)}$ the coordinates, supposedly known, of the ellipsoid centers in the world coordinate frame $\mathscr{R}_w$. $\mathbf{\Delta}_c^{(i)}$ and $K_2^{(i)}$ are calculated from ${}^{w}R_c(\mathbf{\Theta}_f)$ as explained in section III-A. Then $\mathbf{E}_w$, the camera position in $\mathscr{R}_w$, is given by the following weighted average:

$$\mathbf{E}_w = \frac{\sum_{i=1}^{N} \alpha_i ({}^{w}R_c(\mathbf{\Theta}_f)\mathbf{\Delta}_c^{(i)} + \mathbf{C}_w^{(i)})}{\sum_{i=1}^{N} \alpha_i}$$

where

$$\alpha_i = \frac{1}{s(\sqrt{vec(K_2^{(i)})})}$$

$s(.)$ referring to the standard deviation of vector's elements.

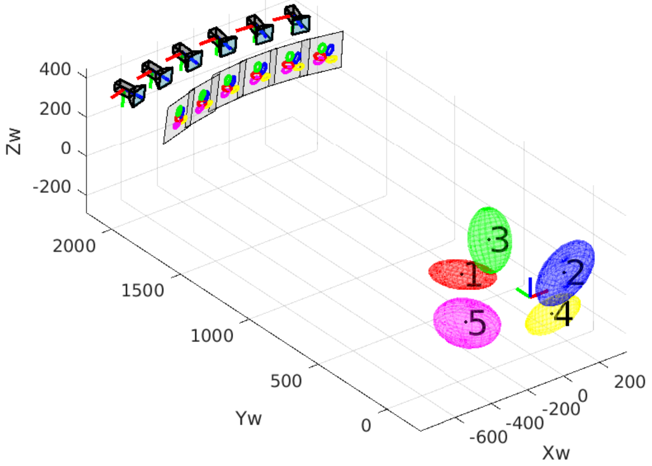


Fig. 4. Simulated scene composed of five ellipsoids and six cameras, whose virtual image planes containing projected ellipses are represented.

IV. Results

A. Simulated Environment Experiments

We first evaluated the performance of our cost function minimization on a generated 3D environment, this in comparison with two other errors from the state of the art. More precisely, the QuadricSLAM method [16] iteratively minimizes a geometric error defined as the distance between the bounding boxes of detected and reprojected ellipses, depending on the six pose parameters, while the analytical solutions presented in [14], [19] are based on an algebraic distance between the vectors formed by the 5 parameters characterizing ellipses in homogeneous coordinates. Note that a closed form solution for pose can be computed when a large number (≥ 12) of correspondences are available [25]. Otherwise, minimization must be done iteratively. In our evaluation, we used these two errors, referred as *QuadricSLAM* and *Algebr. err.* in Fig. 5, as cost functions of the optimization process, to compare them with ours (f'_{discr}, referred as *Discr.*). It is important to note that minimizing *QuadricSLAM* and *Algebr. err.* requires the knowledge of a prior on the camera position, as opposed to our method which requires only an approximate orientation, easier to obtain in practice (gyrometer, computation from vanishing points, etc).

The generated scene (shown in Fig. 4) is composed of five ellipsoids of radii $(18, 12, 6\text{cm})$, $(20, 10, 8\text{cm})$, $(10, 5, 15\text{cm})$, $(2, 9, 15\text{cm})$ and $(15, 12, 10\text{cm})$, arbitrarily placed and oriented, and six cameras positioned approximately at equal distance (average: 2.1m) of the centroid of the ellipsoids. We have successively considered a scene composed of all ellipsoids and only of the first two ones, and have implemented two different magnitudes for initial perturbations on the camera pose (maximum 2° on each Euler angle of the orientation and 5% of the distance between the cameras and the scene on each position, then 10° and 30%). We have also introduced several levels of perturbations on the projected ellipses. To do that, we have added a noise lower than 1, 3, 5, or 7 pixels on the x-y coordinates of six points regularly

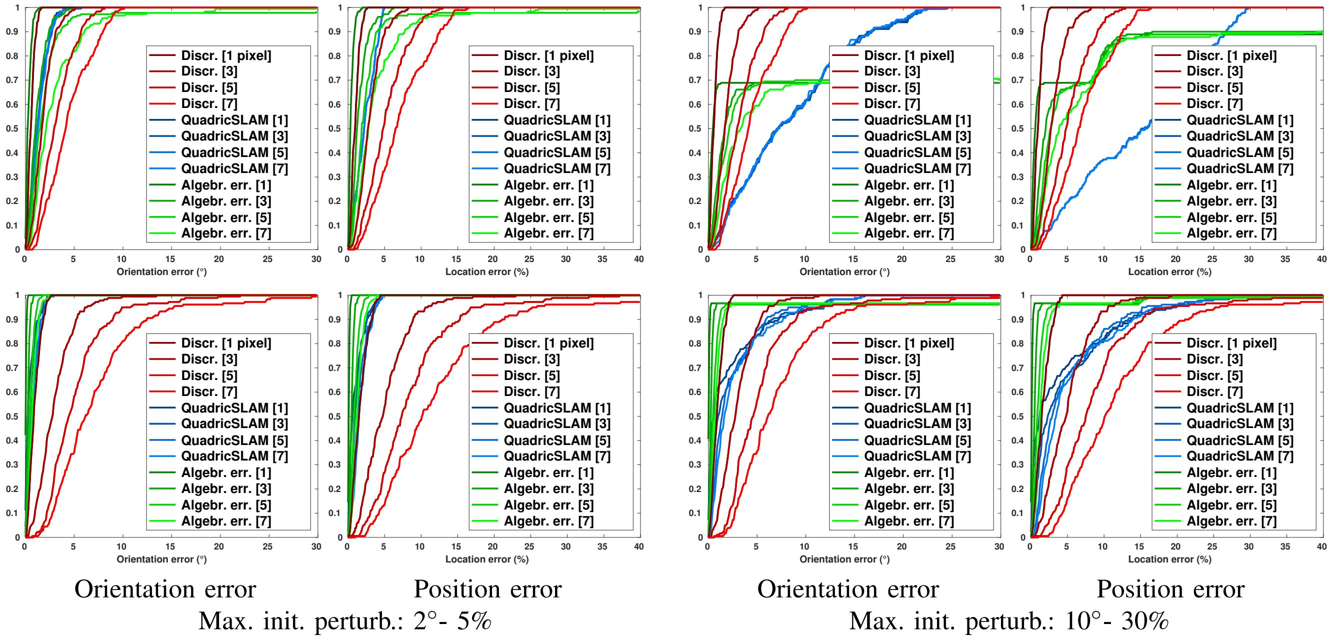


Fig. 5. Simulated environment: cumulative density functions of orientation and position errors, depending on the perturbations on ellipses (1, 3, 5, or 7 pixels), on the maximal initial perturbations (columns), and on the number of ellipsoids in the scene (rows). First row: scene made of 2 ellipsoids. Second row: 5 ellipsoids.

placed on each ground truth ellipse, then we have replaced the original ellipses by those which were interpolated over the six noisy points.

Results presented in Fig. 5 show that our method, which only requires an orientation prior, is robust to the magnitude of the initial perturbation. Indeed, the final errors on camera poses are not affected by the initialisation. However, since discriminant computations include differentiations, our method is sensitive to the noise on ellipses as well as to the number of ellipses detected in the image. By comparison, QuadricSLAM, which requires a whole pose prior, is robust to detection noise, but is sensitive to the initial perturbation. Moreover, when the number of objects is too small, the optimization process does not improve the accuracy of the camera position. Finally, the algebraic distance on ellipses, which also requires a prior for each of the six extrinsic parameters, coherently suffers from detection noise and takes benefit of the number of ellipsoids. In addition, we note that its results highly depend on the initialization accuracy, since a coarse prior can lead to a substantial number of degenerate final poses.

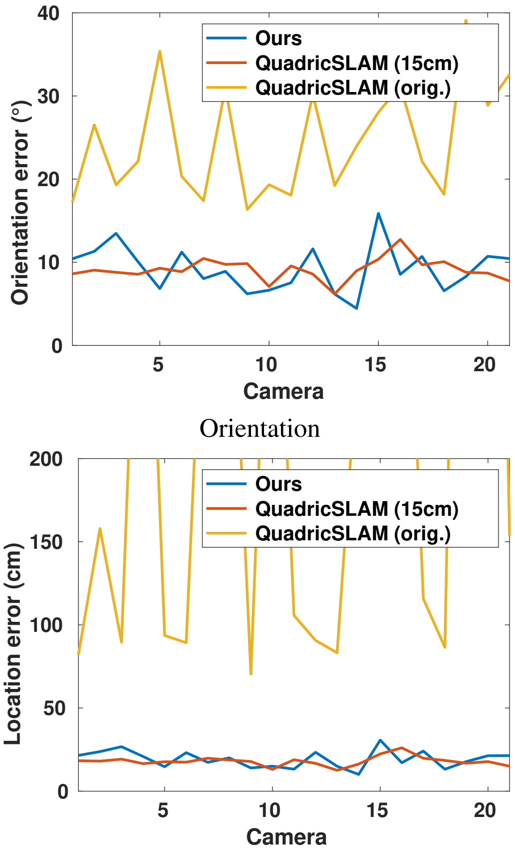
B. *TUM RGB-D Experiments*

For an in-depth evaluation of our method, we have tested it on the publicly available TUM RGB-D Dataset [26] (sequence *Fr2/Desk*). Objects were first detected using YOLOv3 [13], then virtual ellipses were fitted to the bounding boxes, as suggested in [14], [19]. The model was then built using [19]. Such ellipses are necessarily aligned with image axes regardless of the real object orientations, potentially leading to major detection errors. For that reason, and since the critical ellipse detection step is not in the scope of that paper, we have considered only a small subset of 25 cameras and 4 objects of interest (monitor, keyboard, mouse, and cup), such that their virtual ellipses were reasonably fitted to the objects. Four of these images were used for model building, and the remaining 21 for testing.

Robust pose estimation. In practical cases, object detection methods can suffer from false or noisy outputs, and multiple objects identified by the same label cannot be associated with the right correponding 3D ellipsoids. In our experiments, only correct and unambiguous detections are considered. However, the noise issue remains. For this reason, and since we considered only a small number of objects, we have designed a robust method which consists in considering successively all the subsets of 2 or more virtual ellipsoids in the scene (and their corresponding ellipses), then applying the minimization of f'_{discr} to them. The best pose is finally chosen as the one that minimizes the sum of all the distances between the 4 vertices (endpoints of radii) of each reprojected ellipse and their closest points on the corresponding detected ellipse.

For comparison purposes, we have also measured the final pose error obtained by minimizing the QuadricSLAM geometric error. As mentioned earlier, this method requires accurate odometry measurements as initialization. By contrast, our method only requires a coarse prior on its orientation (which can be computed from vanishing points for instance). In our tests, we have compared the performance of both methods, using an initial orientation defined by Euler angles affected by a noise of at least 5° and at most 10° (uniform distribution between these two values), and an initial position for QuadricSLAM affected by a noise of 15 cm. We have also considered the case where no piece of information is given to QuadricSLAM about the camera position (therefore chosen as the origin of the world coordinate frame, located roughly at 2m from the ground truth cameras).

Results presented in Fig. 6 show that our method achieve the same level of performance as QuadricSLAM with accurate initial position (15 cm), but from less and more easily accessible information (coarse prior on camera orientation, instead of accurate priors on both orientation and position). Indeed, the results demonstrate that the QuadricSLAM method cannot handle an arbitrarily initialized camera position.



Orientation

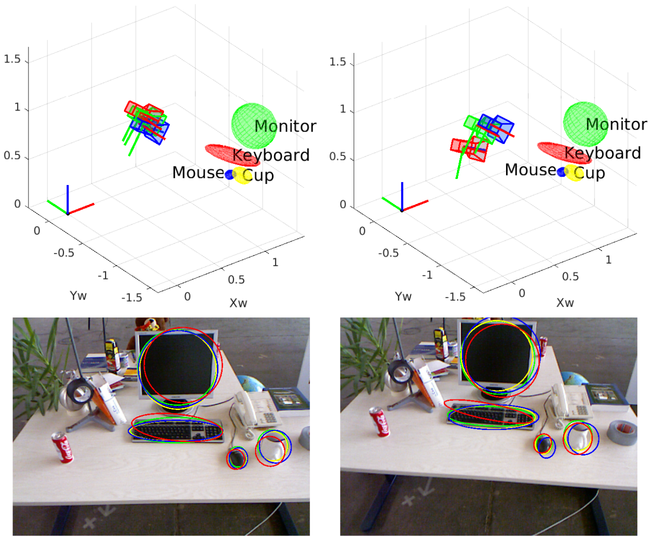
Position

Fig. 6. Final pose errors on TUM RGB-D Dataset with our method and QuadricSLAM, starting from noisy estimations (between 5° and 10° per Euler angle for orientation (both methods), and 15 cm in a random direction for position (only for QuadricSLAM), then the origin of the world coordinate frame (QuadricSLAM)). 10 experiments were conducted per camera, and average results are displayed.

Some visualizations of the pose estimates and reprojected and detected ellipses are presented in Fig. 7. In these examples, ellipses reprojected using our camera pose estimates (blue) are closer to detected ellipses (yellow) than those of QuadricSLAM (red).

V. CONCLUSIONS

We have introduced a novel camera pose estimation method designed to deal with a semantic 3D model made of virtual ellipsoids that correspond to objects of interest present in the scene. Our system considers detected virtual ellipses to model projected objects, then is able to compute the pose from 2 or more correspondences between 2D and 3D entities.



Camera #14

Camera #18

Fig. 7. Illustrations of the TUM RGB-D Dataset experiments. First row: 3D scene model with groundtruth and estimated cameras (Green: ground truth, Blue: ours, Red: QuadricSLAM with accurate position initialization). Second row: images with detected and reprojected ellipses (Yellow: detected ellipses, Green: reprojected with groundtruth camera, Blue: reprojected with our estimated camera, Red: reprojected with accurately initialized QuadricSLAM camera).

Our article has introduced some theoretical advances about pose estimation from correspondences between ellipses and ellipsoids. Indeed, we have shown that the problem consists only in finding the camera orientation and we have then identified the only position compatible with a given correct orientation.

The key features of our algorithm are its ability to solve the kidnapping problem using only a coarse prior on the camera orientation, and its ability to operate on a small number of high-level semantic features robust to image perturbations and low resolutions. The obtained pose can eventually be refined using any classical tracking methods.

APPENDIX 1: $A\mathbf{\Delta} = \sigma B'\mathbf{\Delta}$

Right-multiplying (1) by $\mathbf{\Delta}$ is like writing

$$B\mathbf{\Delta} = \sigma B'\mathbf{\Delta}$$

However, $\mathbf{\Delta}^\top A\mathbf{\Delta}$ is a scalar, thus we have

$$\begin{aligned} B\mathbf{\Delta} &= (A\mathbf{\Delta}\mathbf{\Delta}^\top A - (\mathbf{\Delta}^\top A\mathbf{\Delta} - 1)A)\mathbf{\Delta} \\ &= A\mathbf{\Delta}(\mathbf{\Delta}^\top A\mathbf{\Delta}) - \mathbf{\Delta}^\top A\mathbf{\Delta}A\mathbf{\Delta} + A\mathbf{\Delta} \\ &= (\mathbf{\Delta}^\top A\mathbf{\Delta})A\mathbf{\Delta} - \mathbf{\Delta}^\top A\mathbf{\Delta}A\mathbf{\Delta} + A\mathbf{\Delta} \\ &= A\mathbf{\Delta} \end{aligned}$$

Finally, $A\mathbf{\Delta} = \sigma B'\mathbf{\Delta}$.

APPENDIX 2: $Q(B'^{-1}A) = 0$

Replacing (2) into (1), we obtain:

$$\sigma^2 B'\mathbf{\Delta}\mathbf{\Delta}^\top B' - (\sigma\mathbf{\Delta}^\top B'\mathbf{\Delta} - 1)A = \sigma B'$$

We can then deduce the following expression for A:

$$A = \frac{\sigma}{1 - \sigma\mathbf{\Delta}^\top B'\mathbf{\Delta}}(B' - \sigma B'\mathbf{\Delta}\mathbf{\Delta}^\top B')$$

Thus, denoting I the identity matrix and defining $f = \frac{\sigma}{1-\sigma\mathbf{\Delta}^\top B'\mathbf{\Delta}}$, then left-multiplying by B'^{-1}, we obtain

$$B'^{-1}A = f(I - \sigma\mathbf{\Delta}\mathbf{\Delta}^\top B')$$

Squaring that expression leads to

$$\begin{aligned}(B'^{-1}A)^2 &= f^2(I - \sigma\mathbf{\Delta}\mathbf{\Delta}^\top B')^2 \\ &= f^2(I - 2\sigma\mathbf{\Delta}\mathbf{\Delta}^\top B' + \sigma^2\mathbf{\Delta}(\mathbf{\Delta}^\top B'\mathbf{\Delta})\mathbf{\Delta}^\top B') \\ &= f^2(I - 2\sigma\mathbf{\Delta}\mathbf{\Delta}^\top B' + \sigma^2(\mathbf{\Delta}^\top B'\mathbf{\Delta})\mathbf{\Delta}\mathbf{\Delta}^\top B') \\ &= f^2(I - \sigma(2 - \sigma\mathbf{\Delta}^\top B'\mathbf{\Delta})\mathbf{\Delta}\mathbf{\Delta}^\top B')\end{aligned}$$

Defining $\mu = 1 - \sigma\mathbf{\Delta}^\top B'\mathbf{\Delta} = 1 - \mathbf{\Delta}^\top A\mathbf{\Delta}$:

$$\begin{aligned}(B'^{-1}A)^2 &= f^2(I - \sigma(\mu+1)\mathbf{\Delta}\mathbf{\Delta}^\top B') \\ &= f^2((\mu+1)(I - \sigma\mathbf{\Delta}\mathbf{\Delta}^\top B') - \mu I) \\ &= f(\mu+1)B'^{-1}A - f^2\mu I \\ &= \frac{\sigma}{\mu}(\mu+1)B'^{-1}A - \frac{\sigma^2}{\mu}I\end{aligned}$$

Finally, we have

$$\mu(B'^{-1}A)^2 = \sigma(\mu+1)B'^{-1}A - \sigma^2 I$$

Thus, denoting $Q(x) = \mu x^2 - (\mu+1)\sigma x + \sigma^2$,

$$Q(B'^{-1}A) = 0$$

APPENDIX 3: Link between the roots of a cubic polynomial and the sign of its discriminant

We consider $P(x) = ax^3 + bx^2 + cx + d$, where $(a, b, c, d) \in \mathbb{R}^4$. The discriminant of P is given by:

$$D = 18abcd - 4b^3d + b^2c^2 - 4ac^3 - 27a^2d^2$$

The link between the roots of P and the sign of D is:

- $D > 0$: P has three distinct real roots,
- $D = 0$: P has one or two distinct real roots,
- $D < 0$: P has three disctinct roots, including one real and two complex conjugates.

References

[1] E. Marchand, H. Uchiyama, and F. Spindler, "Pose estimation for augmented reality: A hands-on survey," *IEEE Transactions on Visualization and Computer Graphics*, vol. 22, no. 12, pp. 2633–2651, Dec 2016.

[2] L. Kneip, D. Scaramuzza, and R. Siegwart, "A novel parametrization of the perspective-three-point problem for a direct computation of absolute camera position and orientation," in *CVPR*, 2011.

[3] V. Lepetit, F. Moreno-Noguer, and P. Fua, "Ep*n*p: An accurate $O(n)$ solution to the p*n*p problem," *International Journal of Computer Vision*, vol. 81, no. 2, pp. 155–166, 2009.

[4] R. I. Hartley and A. Zisserman, *Multiple View Geometry in Computer Vision*, 2nd ed. Cambridge University Press, 2004.

[5] C. Xu, L. Zhang, L. Cheng, and R. Koch, "Pose estimation from line correspondences: A complete analysis and a series of solutions," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, pp. 1209–1222, Jun 2017.

[6] T. Taketomi, H. Uchiyama, and S. Ikeda, "Visual slam algorithms: a survey from 2010 to 2016," *IPSJ Transactions on Computer Vision and Applications*, vol. 9, no. 1, p. 16, Jun 2017.

[7] R. B. Girshick, "Fast R-CNN," in *ICCV*, 2015.

[8] S. Ren, K. He, R. B. Girshick, and J. Sun, "Faster R-CNN: towards real-time object detection with region proposal networks," in *NIPS*, 2015.

[9] K. He, G. Gkioxari, P. Dollár, and R. B. Girshick, "Mask R-CNN," in *ICCV*, 2017.

[10] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. E. Reed, C. Fu, and A. C. Berg, "SSD: single shot multibox detector," in *ECCV*, 2016.

[11] J. Redmon, S. K. Divvala, R. B. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *CVPR*, 2016.

[12] J. Redmon and A. Farhadi, "YOLO9000: better, faster, stronger," in *CVPR*, 2017.

[13] ——, "Yolov3: An incremental improvement," http://arxiv.org/abs/1804.02767, 2018.

[14] M. Crocco, C. Rubino, and A. Del Bue, "Structure from motion with objects," in *CVPR*, 2016.

[15] P. Gay, V. Bansal, C. Rubino, and A. Del Bue, "Probabilistic structure from motion with objects (psfmo)," in *ICCV*, 2017.

[16] L. Nicholson, M. Milford, and N. Sünderhauf, "Quadricslam: Dual quadrics from object detections as landmarks in object-oriented slam," *IEEE Robotics and Automation Letters*, vol. 4, no. 1, pp. 1–8, Jan 2019.

[17] J. Li, D. Meger, and G. Dudek, "Semantic mapping for view-invariant relocalization," in *ICRA*, 2019.

[18] S. L. Bowman, N. Atanasov, K. Daniilidis, and G. J. Pappas, "Probabilistic data association for semantic SLAM," in *ICRA*, 2017.

[19] C. Rubino, M. Crocco, and A. D. Bue, "3d object localisation from multi-view image detections," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 40, no. 6, pp. 1281–1294, Jun 2018.

[20] D. S. Wokes and P. L. Palmer, "Autonomous pose determination of a passive target through spheroid modelling," in *AIAA Guidance, Navigation and Control Conference and Exhibit*, Aug 2008.

[21] ——, "Perspective reconstruction of a spheroid from an image plane ellipse," *International Journal of Computer Vision*, vol. 90, no. 3, pp. 369–379, 2010.

[22] D. Eberly, "Reconstructing an ellipsoid from its perspective projection onto a plane," https://www.geometrictools.com/, May 2007, updated version: March 1, 2008.

[23] G. H. Golub and C. F. Van Loan, *Matrix Computations*, 3rd ed. The Johns Hopkins University Press, 1996.

[24] H. Avron, E. Ng, and S. Toledo, "A generalized courant-fischer minimax theorem," Aug 2008. [Online]. Available: https://escholarship.org/uc/item/4gb4t762

[25] V. Gaudillière, G. Simon, and M.-O. Berger, "Perspective-12-Quadric: An analytical solution to the camera pose estimation problem from conic - quadric correspondences," Mar 2019. [Online]. Available: https://hal.archives-ouvertes.fr/hal-02054882

[26] J. Sturm, N. Engelhard, F. Endres, W. Burgard, and D. Cremers, "A benchmark for the evaluation of rgb-d slam systems," in *IROS*, 2012.