3D Object Localisation from Multi-view Image Detections

Cosimo Rubino, *Student Member, IEEE,* Marco Crocco, *Member, IEEE,* Alessio Del Bue, *Member, IEEE,*

Abstract—In this work we present a novel approach to recover objects 3D position and occupancy in a generic scene using only 2D object detections from multiple view images. The method reformulates the problem as the estimation of a quadric (ellipsoid) in 3D given a set of 2D ellipses fitted to the object detection bounding boxes in multiple views. We show that a closed-form solution exists in the dual-space using a minimum of three views while a solution with two views is possible through the use of non-linear optimisation and object constraints on the size of the object shape. In order to make the solution robust toward inaccurate bounding boxes, a likely occurrence in object detection methods, we introduce a data preconditioning technique and a non-linear refinement of the closed form solution based on implicit subspace constraints. Results on synthetic tests and on different real datasets, involving challenging scenarios, demonstrate the applicability and potential of our method in several realistic scenarios.

Index Terms—Multi-view geometry, 3D localisation, object detection, conics optimisation

1 Introduction

The detection and localisation of objects in a generic scene is a fundamental step for the understanding of visual world, which can enable higher level semantic tasks and disruptive applications in every day life. Although the localisation of objects has been restricted mostly onto the image plane, several attempts have been tried to lift such reasoning into 3D by instantiating the object detection problem as a 3D object localisation problem.

This reasoning in 3D has been generally treated in two distinct frameworks. First with a learning paradigm, where the object-based classification output provides directly the object position and its 3D orientation [1], [2], [3], [4], [5], [6], [7], [8]. While this 6D pose problem was initially studied for custom object classes (e.g. a head [9]), now the recent trend is to provide general solutions for several classes of objects [10], [11]. However these approaches often require the construction of custom pose dependent datasets that can be complex to achieve.

A second approach consists in including geometrical reasoning explicitly in the object detection framework. Recent works have clearly pointed out that bridging the gap between object detection and multi-view geometry might provide surprising improvements in classical approaches. Starting from the work of Hoeim et al. [12], the inclusion of 3D scene reasoning and rules has provided higher detection accuracies, up to the estimation of coarse 3D geometry from single views [13]. Notably, attempts of unifying geometry and object representation have been achieved by defining an elaborated Maximum a Posteriori (MAP) inferences [14] or bundle adjustment with objects [15]. This way of pursuing high-level object reasoning in multi-view geometry has also inspired novel methods in Simultaneous Localisation and Mapping (SLAM) [16]. Semantic information has been used, on the other hand, to infer jointly the 3D shape and cameras viewpoints [10].

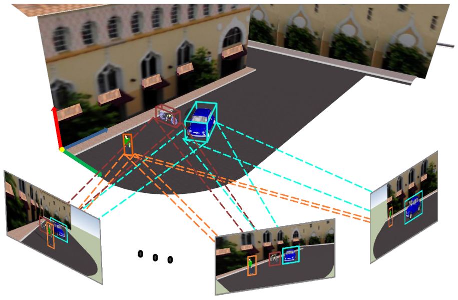


Fig. 1. Example of a set of images of a given 3D rigid scene taken from a camera at different viewpoints. The problem consists in recovering the 3D occupancy of each object given the 2D bounding boxes detected at each image frame.

Our work takes a different path from previous methods by showing that it is possible to obtain accurate 3D object localisations and occupancy using mainly multi-view geometric relations given 2D bounding boxes only. Starting from the location and size of a set of bounding boxes detected in a generic sequence (see Fig. 1 for a graphical representation), we show that the localisation of objects in 3D can be instantiated as a conics optimisation problem with a closed-form solution in the dual space. Our solution does not use directly bounding boxes (which is a piecewise-defined curve), but we show that the problem is mathematically feasible given a set of ellipses fitted to the original 2D bounding boxes as extracted by the detectors. In this way, we can reformulate the problem as the estimation of 3D quadrics given known camera matrices and a set of 2D conics in multiple views. We finally show that, for the ill-posed case of two views, it is possible to apply a further non-linear optimisation with object based constraints, obtaining a performance close to the well-posed solution. Moreover,

- *C. Rubino, M. Crocco, and A. Del Bue are with the Visual Geometry and Modelling (VGM) Lab, Istituto Italiano di Tecnologia, Via Morego 30, 16163 Genova, Italy.*
E-mail: alessio.delbue@iit.it

non-linear optimisation, as well as data preconditioning, in general boosts results whenever the bounding boxes are inaccurate as it might happen in realistic scenarios. In such a way, improved results are achieved in challenging real scenarios on different freely available datasets. Differently from previous works, we do not need geometric or semantic priors, apart from 2D bounding boxes from detections, nor advanced detectors yielding the 3D position of the objects. Even without this information, the proposed approach is in general quicker and more accurate than current methods, which make use of stronger semantics.

The rest of the paper is structured as follows. Section 2 provides background on related work while Section 3 defines our problem and the associated mathematical formalisation. Section 4 presents the proposed closed-form solution together with the non-linear optimisation approach in Section 5. Experiments on synthetic and real data are discussed in Section 6 and 7 respectively. Finally, Section 8 presents concluding remarks about the proposed approach.

2 Related Work

In the last years many approaches were developed to infer the location and orientation of objects from images in a general 3D scenario. These works are mainly categorised by the type of constraints assumed during inference, given either by the object classes, the complexity of the scene (indoor, outdoor) and the information available (single image, video or RGBD). In this review we will restrict to single or multiple views methods, for which our approach is more closely related, and neglect the RGBD case.

As the most challenging scenario, strong efforts have been devoted to the study of single image pose estimation problems. One of the first examples in the literature [17] defined image heuristics for the estimation of position and orientation of piecewise planar objects (e.g. a chair). The so-called Origami world assumption was dealing with specific classes, but yet was able to provide reasonable heuristics to deal with the ill-posedness of recovering the object pose from a single view. These pre-defined geometrical heuristics were the standard approach for these early works [18] but quickly revealed their limitation in modelling the complexity of the real word geometrical composition.

This led to the need to learn image to object relations in order to generalise pose estimation in 3D to several classes of objects. In many cases a training phase is performed using images of a specific category of objects from different viewpoints. This severely underconstrained problem has to be solved by considering strong semantic information about each object and the context, making often the algorithms specific to a class or a subclass of objects. Many works have exploited 3D object models (in specific CAD wireframes) to get a 3D interpretation of the scene. Zia et al. [19] [20] [21] used the CAD models of cars to reconstruct the scene and the objects, including additional information about the ground plane. Pepik et al. [3] reformulated the model as a 3D deformable part model by learning the part appearances according to the CAD model. Liebelt and Schmid [22] trained a multi-view detector to learn the object appearance, and then linked the geometric information to the 2D training data to perform the pose estimation for generic object classes.

When multiple images are available, recent works have tried to include geometrical reasoning to explicitly use constraints given by the multiple views. Bao et al. [23] tried to deduce both the viewpoint motion among multiple images and the pose of the objects using a part-based object detector, improving the performances through a dense reconstruction and by incorporating semantic information as category-level shape priors (learned from the object) [24]. To reach the same goal a monocular SLAM approach was used by Dame et al. [25], combining it with shape priors-based 3D tracking and 3D reconstruction approaches, while Findler et al. [26] reduced all the objects to 3D bounding boxes with each side being a planar approximation of the object: in this case the localisation and orientation are estimated considering the perspective reprojections of the 3D bounding box faces onto the image frames.

Differently from these methods, the proposed approach uses the 2D bounding boxes to define, after ellipse fitting, a quadric reconstruction problem from multiple conics in 2D. Previous approaches have studied the problem of quadrics estimation for the cases of three [27] and two views [28], mainly related to obtain a reconstruction from object contours. Ma et al. [27] described all the steps to perform the reconstruction from three views given object contours, showing that two views only are not enough to recover a generic ellipsoid. Cross et al. [28] illustrated a two views method to reconstruct quadrics using again object outlines extracted and matched from multiple images. This solution was possible by using additional epipolar constraints between matched points in the images. Differently from these works, we deal with the case of multiple objects in multiple frames and we provide pre-conditioning and non-linear optimisation in order to solve for the problem under perspective projection.

The work of Crocco et al. [29] presented a solution to the quadrics reconstruction problem with the simpler orthographic camera model, which can not be used with the experimental scenarios showing strong perspective effects, as the ones presented in this work. Moreover, our approach is resilient if some of the detections are missing, differently from [29] that solves the problem using the factorisation of a complete matrix containing the ellipses parameterisation for every object at every frame. Lastly, [29] does not take into account any regularisation method as we do in this work, which improves considerably results when inaccurate bounding box estimates are present.

3 Problem statement

Let us consider a set of image frames $f = 1 \dots F$ representing a 3D scene under different viewpoints. A set of $i = 1 \dots N$ rigid objects is placed in arbitrary position and each object can be detected in each of the F images. Each object i in each image frame f is identified by a 2D bounding box B_{if}, given by a generic object detector. The bounding box is defined by a triplet of parameters $B_{if} = \{w_{if}, h_{if}, \mathbf{b}_{if}\}$, where w_{if} and h_{if} are two scalars for the bounding box height and width respectively, while $\mathbf{b}_{if}$ is a 2-vector defining the bounding box centre.

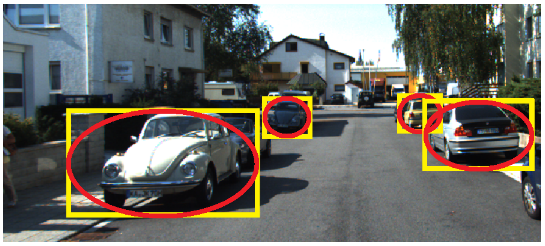


Fig. 2. Example of Bounding Boxes (yellow) and corresponding fitted ellipses (red) for a set of objects.

Our goal is to estimate the position and space occupancy of each object in the 3D scene given the 2D bounding boxes and by using multi-view constraints. In order to ease the mathematical formalisation of the problem, we move from a bounding box representation of an object to an ellipsoid one. This step is performed by associating at each B_{if} an ellipse $\hat{\mathtt{C}}_{if}$ that inscribes the bounding box, as shown in Fig. 2. In particular, each ellipse is centred in $\mathbf{b}_{if}$ and is aligned to the image axes, with axes length equal to w_{if} and h_{if}. The aim of our problem is to find the 3D ellipsoids $\mathtt{Q}_i$ whose projections onto the image planes best fit the 2D ellipses $\hat{\mathtt{C}}_{if}$. This will solve for both the 3D localisation and occupancy of each object starting from the image detections in the different views.

In the following, we represent each ellipse using the homogeneous quadratic form of a conic equation:

$$\mathbf{u}^\top \hat{\mathtt{C}}_{if}\ \mathbf{u} = 0, \tag{1}$$

where $\mathbf{u} \in \mathbb{R}^3$ is the homogeneous vector of a generic 2D point belonging to the conic defined by the symmetric matrix $\hat{\mathtt{C}}_{if} \in \mathbb{R}^{3\times3}$. The conic has five degrees of freedom given by the six elements of the lower triangular part of the symmetric matrix $\hat{\mathtt{C}}_{if}$, except one for the scale since Eq. (1) is homogeneous in $\mathbf{u}$ [30]. Similarly to the ellipses, we represent the ellipsoids in the 3D space with the homogeneous quadratic form of a quadric equation:

$$\mathbf{x}^\top \mathtt{Q}_i\, \mathbf{x} = 0, \tag{2}$$

where $\mathbf{x} \in \mathbb{R}^4$ is an homogeneous 3D point belonging to the quadric defined by the symmetric matrix $\mathtt{Q}_i \in \mathbb{R}^{4\times4}$. The quadric has nine degrees of freedom, given by the ten elements of the symmetric matrix $\mathtt{Q}_i$ up to one for the overall scale.

Each quadric $\mathtt{Q}_i$, when projected onto the image plane, gives a conic denoted by $\mathtt{C}_{if} \in \mathbb{R}^{3\times3}$. The relationship between $\mathtt{Q}_i$ and $\mathtt{C}_{if}$ is defined by the projection matrices $\mathtt{P}_f = \mathtt{K}_f[\mathtt{R}_f|\mathbf{t}_f] \in \mathbb{R}^{3\times4}$, where $\mathtt{K}_f \in \mathbb{R}^{3\times3}$, $\mathtt{R}_f \in \mathbb{R}^{3\times3}$ and $\mathbf{t}_f \in \mathbb{R}^{3\times1}$ are respectively the matrix of the intrinsic parameters, the rotation matrix and the translation vector of the cameras associated to each frame f. Such matrices are assumed to be known (i.e. the camera is calibrated) and can be estimated from the image sequence using standard self-calibration methods.

4 Dual space fitting

Since the relationship between $\mathtt{Q}_i$ and $\mathtt{C}_{if}$ is not straightforward in the primal space, i.e. the Euclidean space of 3D points (2D points in the images), it is convenient to reformulate it in dual space, i.e. the space of the planes (lines in the images) [27], [28]. In particular, the conics in 2D can be represented by the envelope of all the lines tangent to the conic curve, while the quadrics in 3D can be represented by the envelope of all the planes tangent to the quadric surface. Hence, the dual quadric is defined by the matrix $\mathtt{Q}_i^* = adj(\mathtt{Q}_i)$, where adj is the adjoint operator, and the dual conic is defined by $\mathtt{C}_{if}^* = adj(\mathtt{C}_{if})$ [30]. Considering that the dual conic $\mathtt{C}_{if}^*$, like the primal one, is defined up to an overall scale factor β_{if}, the relationship between a dual quadric and its dual conic projections $\mathtt{C}_{if}^*$ can be written as:

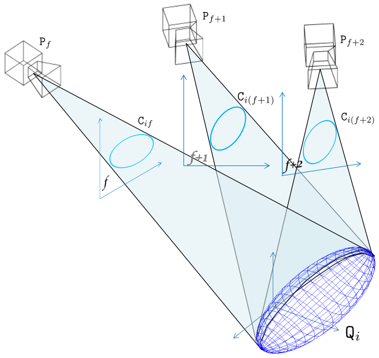


Fig. 3. Example of a set of conics $\mathtt{C}_{if}$, $\mathtt{C}_{i(f+1)}$ and $\mathtt{C}_{i(f+2)}$ which represents the outlines in 3 frames of a given quadric $\mathtt{Q}_i$.

$$\beta_{if}\mathtt{C}_{if}^* = \mathtt{P}_f \mathtt{Q}_i^* \mathtt{P}_f^\top. \tag{3}$$

In order to recover $\mathtt{Q}_i^*$ in closed form from the set of dual conics $\{\mathtt{C}_{if}^*\}_{f=1\ldots F}$, we have to re-arrange Eq. (3) into a linear system. Let us define $\mathbf{v}_i^* = vech(\mathtt{Q}_i^*)$ and $\mathbf{c}_{if}^* = vech(\mathtt{C}_{if}^*)$ as the vectorisation of symmetric matrices $\mathtt{Q}_i^*$ and $\mathtt{C}_{if}^*$ respectively[1]. Then, let us arrange the products of the elements of $\mathtt{P}_f$ and $\mathtt{P}_f^\top$ in a single matrix $\mathtt{G}_f \in \mathbb{R}^{6\times10}$ as follows [31]:

$$\mathtt{G}_f = \mathtt{D}(\mathtt{P} \otimes \mathtt{P})\mathtt{E} \tag{4}$$

where $\otimes$ is the Kronecker product and matrices $\mathtt{D} \in \mathbb{R}^{6\times9}$ and $\mathtt{E} \in \mathbb{R}^{16\times10}$ are two matrices such that $vech(\mathtt{X}) = \mathtt{D}\ vec(\mathtt{X})$ and $vec(\mathtt{Y}) = \mathtt{E}\ vech(\mathtt{Y})$ respectively, where $\mathtt{X} \in \mathbb{R}^{9\times9}$ and $\mathtt{Y} \in \mathbb{R}^{16\times16}$ are two symmetric matrices[2]. The structure of the matrix $\mathtt{G}_f$ is made explicit in Eq. (4), where p_{qr} are all the entries of the matrix $\mathtt{P}$, for $q = 1\ldots3$ and $r = 1\ldots4$. Given $\mathtt{G}_f$, we can rewrite Eq. (3) as:

$$\beta_{if}\mathbf{c}_{if}^* = \mathtt{G}_f \mathbf{v}_i^*. \tag{6}$$

4.1 Direct solution for ellipsoid reconstruction

In order to get a unique solution for $\mathbf{v}_i^*$ at least three image frames are needed. Therefore, stacking column-wise Eq. (6) for $f = 1\ldots F$, with $F \geq 3$, we obtain:

$$\mathtt{M}_i\mathbf{w}_i = \mathbf{0}_{6F}, \tag{7}$$

[1] The operator $vech$ serialises the elements of the lower triangular part of a symmetric matrix, such that, given a symmetric matrix $\mathtt{X} \in \mathbb{R}^{n\times n}$, the vector $\mathbf{x}$, defined as $\mathbf{x} = vech(\mathtt{X})$, is $\mathbf{x} \in \mathbb{R}^g$ with $g = \frac{n(n+1)}{2}$.

[2] The operator vec serialises all the elements of a generic matrix.

$$\mathtt{G}_f = \begin{bmatrix} {p_{11}}^2 & 2\,p_{12}p_{11} & 2p_{13}p_{11} & 2\,p_{14}p_{11} & {p_{12}}^2 & 2p_{13}p_{12} & 2p_{14}p_{12} & {p_{13}}^2 & 2p_{13}p_{14} & {p_{14}}^2 \\ p_{21}p_{11} & p_{21}p_{12}+p_{22}p_{11} & p_{23}p_{11}+p_{21}p_{13} & p_{24}p_{11}+p_{21}p_{14} & p_{22}p_{12} & p_{22}p_{13}+p_{23}p_{12} & p_{22}p_{14}+p_{24}p_{12} & p_{23}p_{13} & p_{23}p_{14}+p_{24}p_{13} & p_{24}p_{14} \\ p_{31}p_{11} & p_{31}p_{12}+p_{32}p_{11} & p_{33}p_{11}+p_{31}p_{13} & p_{34}p_{11}+p_{31}p_{14} & p_{32}p_{12} & p_{32}p_{13}+p_{33}p_{12} & p_{32}p_{14}+p_{34}p_{12} & p_{33}p_{13} & p_{33}p_{14}+p_{34}p_{13} & p_{34}p_{14} \\ {p_{21}}^2 & 2p_{22}p_{21} & 2p_{23}p_{21} & 2p_{24}p_{21} & {p_{22}}^2 & 2p_{23}p_{22} & 2p_{24}p_{22} & {p_{23}}^2 & 2p_{23}p_{24} & {p_{24}}^2 \\ p_{31}p_{21} & p_{31}p_{22}+p_{32}p_{21} & p_{33}p_{21}+p_{31}p_{23} & p_{34}p_{21}+p_{31}p_{24} & p_{32}p_{22} & p_{32}p_{23}+p_{33}p_{22} & p_{32}p_{24}+p_{34}p_{22} & p_{33}p_{23} & p_{33}p_{24}+p_{34}p23 & p_{34}p_{24} \\ {p_{31}}^2 & 2p_{32}p_{31} & 2\,p_{33}p_{31} & 2\,p_{34}p_{31} & {p_{32}}^2 & 2p_{33}p_{32} & 2\,p_{34}p_{32} & {p_{33}}^2 & 2p_{33}p_{34} & {p_{34}}^2 \end{bmatrix} \quad (5)$$

where $\mathbf{0}_x$ denotes a column vector of zeros of length x, and the matrix $\mathtt{M}_i \in \mathbb{R}^{6F\times(10+F)}$ and the vector $\mathbf{w}_i \in \mathbb{R}^{10+F}$ are defined as follows:

$$\mathtt{M}_i = \begin{bmatrix} \mathtt{G}_1 & \text{-}\mathbf{c}^*_{i1} & \mathbf{0}_6 & \mathbf{0}_6 & \dots & \mathbf{0}_6 \\ \mathtt{G}_2 & \mathbf{0}_6 & \text{-}\mathbf{c}^*_{i2} & \mathbf{0}_6 & \dots & \mathbf{0}_6 \\ \mathtt{G}_3 & \mathbf{0}_6 & \mathbf{0}_6 & \text{-}\mathbf{c}^*_{i3} & \dots & \mathbf{0}_6 \\ \vdots & \mathbf{0}_6 & \mathbf{0}_6 & \mathbf{0}_6 & \ddots & \mathbf{0}_6 \\ \mathtt{G}_F & \mathbf{0}_6 & \mathbf{0}_6 & \mathbf{0}_6 & \dots & \text{-}\mathbf{c}^*_{iF} \end{bmatrix}, \quad \mathbf{w}_i = \begin{bmatrix} \mathbf{v}^*_i \\ \boldsymbol{\beta}_i \end{bmatrix}, \quad (8)$$

with $\boldsymbol{\beta}_i = [\beta_{i1}, \beta_{i2}, \cdots, \beta_{iF}]^\top$. Note that in real cases the ellipses $\hat{\mathtt{C}}_{if}$ computed by a general purpose object detector might be inaccurate regarding the location of the bounding box and the window sizes. Likewise, this will have an effect on the ellipsoid fitting, inducing an error on the $\mathtt{C}_{if}$. For this reason, if $\tilde{\mathtt{M}}_i$ is the matrix given by object detections, we can find the solution by minimising:

$$\tilde{\mathbf{w}}_i = \arg\min_{\mathbf{w}} \|\tilde{\mathtt{M}}_i \mathbf{w}\|_2^2 \quad s.t. \quad \|\mathbf{w}\|_2^2 = 1, \quad (9)$$

where the equality constraint $\|\mathbf{w}\|_2^2 = 1$ avoids the trivial zero solution. The minimisation problem in Eq. (9) can be solved by applying the SVD to the $\tilde{\mathtt{M}}_i$ matrix, taking the right singular vector associated to the minimum singular value. The first 10 entries of $\tilde{\mathbf{w}}_i$ are the vectorised elements of the estimated dual quadric, denoted by $\tilde{\mathbf{v}}^*_i$. To get back the estimated matrix of the quadric in the primal space, we obtain first the dual estimated quadric by:

$$\tilde{\mathtt{Q}}^*_i = vech^{-1}(\tilde{\mathbf{v}}^*_i) \quad (10)$$

and subsequently apply the following relation:

$$\tilde{\mathtt{Q}}_i = adj^{-1}(\tilde{\mathtt{Q}}^*_i) \quad (11)$$

where adj^{-1} denotes the inverse of the adjoint operator.

4.2 Conics and quadrics pre-conditioning

Possible inaccuracies in the bounding boxes and in the projection matrices estimation, embedded in the matrix $\tilde{\mathtt{M}}_i$, propagate to the solution $\tilde{\mathbf{w}}_i$ in a quite complex manner, as described by the perturbation theory [32]. In general, if the matrix $\tilde{\mathtt{M}}_i$ is ill-conditioned, small errors on its entries may result in a grossly inaccurate solution. One of the main sources of ill-conditioning is the diversity in the magnitude of the entries of $\tilde{\mathtt{M}}_i$.

In order to gain a deeper insight into the source of such diversity, let us express a generic ellipse $\mathtt{C}^*_{if}$ in dual space as an ellipse $\breve{\mathtt{C}}^*_{if}$, centred in the image centre and with normalised axes length, subjected to an homogeneous transformation $\mathtt{H}_{if}$, as follows:

$$\mathtt{C}^*_{if} = \mathtt{H}_{if} \breve{\mathtt{C}}^*_{if} \mathtt{H}^\top_{if}, \quad (12)$$

where:

$$\mathtt{H}_{if} = \begin{bmatrix} h & 0 & t^c_1 \\ 0 & h & t^c_2 \\ 0 & 0 & 1 \end{bmatrix}, \quad \breve{\mathtt{C}}^*_{if} = \begin{bmatrix} c^*_{11} & c^*_{12} & 0 \\ c^*_{12} & c^*_{22} & 0 \\ 0 & 0 & \text{-}1 \end{bmatrix}. \quad (13)$$

In details, t^c_1 and t^c_2 are the coordinates of the ellipse centre and $h = \sqrt{l_1^2 + l_2^2}$, where $l_1, l_2 \in \mathbb{R}$ are the two semi axes of the ellipse. Using Eqs. (12) and (13) we can express the vectorised conic $\mathbf{c}^*_{if}$ as:

$$\mathbf{c}^*_{if} = \begin{bmatrix} h^2 c^*_{11} - {t^c_1}^2 \\ h^2 c^*_{12} - t^c_1 t^c_2 \\ -t^c_1 \\ h^2 c^*_{22} - {t^c_2}^2 \\ -t^c_2 \\ -1 \end{bmatrix}. \quad (14)$$

Here the first, second and fourth element of $\mathbf{c}^*_{if}$ have a quadratic dependence from both the ellipse translation and axes length. When a wide baseline sequence is considered the projection matrices $\mathtt{P}_f$ will project the same quadric on very different conics in terms of size and translation. This in turn will cause sharp diversity in the magnitudes of the vectors $\mathbf{c}^*_{if}$ across different views f, thus yielding an ill-conditioned matrix $\tilde{\mathtt{M}}_{if}$. Moreover the translation and axes term sum up together in the elements of $\mathbf{c}^*_{if}$, despite their different geometrical meaning. In the case of small ellipses far from the image centre the terms related to the ellipse size and shape become negligible with respect to the translation terms, and consequently even small errors on $\mathbf{c}^*_{if}$ affect negatively the reconstruction of the ellipsoid. This is because the information on the ellipsoid shape, embedded in the elements c^*_{11}, c^*_{12} and c^*_{22} do not prevail over the translation errors on t^c_1 and t^c_2.

To cope with these issues, we devised a preconditioning strategy inspired by coordinate normalisation in multi-view geometry. In detail, once ellipses centres and axes have been extracted we apply the inverse of the homogeneous transformation $\mathtt{H}_{if}$ to both the members of Eq. (3), obtaining:

$$\beta_{if} \breve{\mathtt{C}}^*_{if} = \mathtt{H}^{-1}_{if} \mathtt{P}_f \mathtt{Q}^*_i \mathtt{P}^\top_f \mathtt{H}^{-\top}_{if}. \quad (15)$$

Finally we build the matrix $\tilde{\mathtt{M}}_i$ collecting the new conics $\breve{\mathtt{C}}^*_{if}$ and the new projection matrices $\mathtt{H}^{-1}_{if}\mathtt{P}_f$ and substituting them to $\mathtt{C}^*_{if}$ and $\mathtt{P}_f$ respectively. In this way all the new conics are centred to the origin and normalised in size.

The dual quadric representation suffers from the same problem already seen for the dual conic, i.e. if the quadric centre is far from the origin, the terms accounting for the ellipsoid shape become negligible with respect to the terms accounting for its position, possibly causing an inaccurate ellipsoid reconstruction. To cope with this issue, we perform a translation of the 3D coordinates, moving the quadric at the centre of the 3D scene. Obviously the exact quadric is unknown, but we can assume that its translation parameters are well approximated by the quadric $\tilde{\mathtt{Q}}_i^*$ i.e. the result of the optimisation problem in Eq. (9). To perform such translation, we can rewrite Eq. (15) in the following way:

$$\beta_{if}\breve{\mathtt{C}}^*_{if} = \mathtt{H}^{-1}_{if}\mathtt{P}_f\mathtt{T}_i\mathtt{T}_i^{-1}\mathtt{Q}_i^*\mathtt{T}_i^{-\top}\mathtt{T}_i^{\top}\mathtt{P}_f^{\top}\mathtt{H}_{if}^{-\top}\,, \qquad (16)$$

where $\mathtt{T}_i$ is the translation matrix whose last column contains the translation parameters of $\tilde{\mathtt{Q}}_i^*$. Then, we consider the terms $\mathtt{H}^{-1}_{if}\mathtt{P}_f\mathtt{T}_i$ in (16) as the new projection matrices and we build accordingly a new matrix $\tilde{\mathtt{M}}_i^c$ related to the centred quadric. Next, we solve the new problem:

$$\tilde{\mathbf{w}}_i^c = \arg\min_{\mathbf{w}} \|\tilde{\mathtt{M}}_i^c\mathbf{w}\|_2^2 \quad s.t. \quad \|\mathbf{w}\|_2^2 = 1, \qquad (17)$$

and we find the dual quadric $\tilde{\mathtt{Q}}_i^{*c}$ from the vector $\tilde{\mathbf{w}}_i^c$, according to the same procedure used to obtain $\tilde{\mathtt{Q}}_i^*$ from $\tilde{\mathbf{w}}_i$. Finally we translate $\tilde{\mathtt{Q}}_i^{*c}$ to the correct 3D location, applying the translation matrix $\mathtt{T}_i$ to $\tilde{\mathtt{Q}}_i^{*c}$ and obtaining the final solution $\tilde{\mathtt{Q}}_i^{*f}$ as follows:

$$\tilde{\mathtt{Q}}_i^{*f} = \mathtt{T}_i\tilde{\mathtt{Q}}_i^{*c}\mathtt{T}_i^{\top}. \qquad (18)$$

Notice that with such procedure, the translation parameters have been substantially decoupled from the shape parameters of the ellipsoid, thus reducing the ill-conditioning problems arising from the addition of the two kinds of parameters in the entries of the dual quadric. As a final step we obtain the solution in the primal space as the inverse adjoint of $\tilde{\mathtt{Q}}_i^{*f}$.

5 Non-linear optimisation

As seen in the previous section, moving to the dual space allows an efficient linearisation of the problem. However this implies a drawback: the algebraic minimisation is carried on the dual quadrics in Eq. (17) and the primal one is obtained by a matrix inversion as in Eq. (11). Thus, in presence of relevant errors in the bounding boxes estimation, the given primal quadric could lead to solutions that are nearly degenerate ellipsoids or even to a different quadric like an hyperboloid. Notice that this problem is connected to, but different from, the one related to ill-conditioning of the matrix $\tilde{\mathtt{M}}_i$ and has to be faced with a different strategy. To overcome this problem, we devised a non-linear cost function that depends on a new set of variables, in which the dual quadric is forced to lie on the subspace of ellipsoids.

In detail, a generic ellipsoid in dual space $\mathtt{Q}^*$ can be rewritten as follows:

$$\mathtt{Q}^* = \mathtt{Z}\,\breve{\mathtt{Q}}^*\mathtt{Z}^{\top} \qquad (19)$$

where $\breve{\mathtt{Q}}^*$ is an ellipsoid centred on the origin and with the axes aligned to the 3D coordinates and $\mathtt{Z}$ is an homogeneous transformation, accounting for an arbitrary rotation and translation. It turns out that $\mathtt{Z}$ and $\breve{\mathtt{Q}}^*$ can be written respectively as:

$$\mathtt{Z} = \begin{bmatrix} \mathtt{R}(\boldsymbol{\theta}) & \mathbf{t} \\ \mathbf{0}_3^{\top} & 1 \end{bmatrix}, \quad \breve{\mathtt{Q}}^* = \begin{bmatrix} a^2 & 0 & 0 & 0 \\ 0 & b^2 & 0 & 0 \\ 0 & 0 & c^2 & 0 \\ 0 & 0 & 0 & \text{-1} \end{bmatrix} \qquad (20)$$

where $\mathbf{t} = [t_1, t_2, t_3]^{\top}$ is the translation vector, $\mathtt{R}(\boldsymbol{\theta})$ is the rotation matrix function of the Euler angles $\boldsymbol{\theta} = [\theta_1, \theta_2, \theta_3]^{\top}$ and a, b, c are the three semi-axes of the ellipsoid [3]. Therefore, we can express every ellipsoid in terms of the nine parameters $\theta_1, \theta_2, \theta_3, t_1, t_2, t_3, a, b, c$. Now, defining the vector $\mathbf{e} \in \mathbb{R}^{9+F}$ as $\mathbf{e} = [\theta_1, \theta_2, \theta_3, t_1, t_2, t_3, a, b, c, \beta_1, \ldots \beta_F]^{\top}$ we can evaluate a functional form of the vector $\mathbf{w}(\mathbf{e})$ as follows:

$$\mathbf{w}(\mathbf{e}) = \begin{bmatrix} r_{11}(\boldsymbol{\theta})^2a^2 + r_{12}(\boldsymbol{\theta})^2b^2 + r_{13}(\boldsymbol{\theta})^2c^2 - t_1^2 \\ r_{11}(\boldsymbol{\theta})r_{21}(\boldsymbol{\theta})a^2 + r_{12}(\boldsymbol{\theta})r_{22(\boldsymbol{\theta})}b^2 + r_{13}(\boldsymbol{\theta})r_{23}(\boldsymbol{\theta})c^2 - t_1t_2 \\ r_{11}(\boldsymbol{\theta})r_{31}(\boldsymbol{\theta})a^2 + r_{12}(\boldsymbol{\theta})r_{32}(\boldsymbol{\theta})b^2 + r_{13}(\boldsymbol{\theta})r_{33}(\boldsymbol{\theta})c^2 - t_1t_3 \\ -t_1 \\ r_{21}(\boldsymbol{\theta})^2a^2 + r_{22}(\boldsymbol{\theta})^2b^2 + r_{23}(\boldsymbol{\theta})^2c^2 - t_2^2 \\ r_{21}(\boldsymbol{\theta})r_{31}(\boldsymbol{\theta})a^2 + r_{22}(\boldsymbol{\theta})r_{32}(\boldsymbol{\theta})b^2 + r_{23}(\boldsymbol{\theta})r_{33}(\boldsymbol{\theta})c^2 - t_2t_3 \\ -t_2 \\ r_{31}(\boldsymbol{\theta})^2a^2 + r_{32}(\boldsymbol{\theta})^2b^2 + r_{33}(\boldsymbol{\theta})^2c^2 - {t_3}^2 \\ -t_3 \\ -1 \\ \beta_1 \\ \vdots \\ \beta_F \end{bmatrix} \qquad (21)$$

where the terms in $r_{mn} \mid m, n = 1, \ldots, 3$ are the entries of the rotation matrix $\mathtt{R}(\boldsymbol{\theta})$. Hence, the new problem in the variables $\mathbf{e}$ can be reformulated as:

$$\tilde{\mathbf{e}}_i = \arg\min_{\mathbf{e}} \|\tilde{\mathtt{M}}_i^c\mathbf{w}(\mathbf{e})\|_2^2 \qquad (22)$$

Notice that this formulation forces the solution to be an ellipsoid without imposing explicit constraints.

If some prior knowledge on the range of the axes length is available, i.e. when the detected objects have a known scale and aspect ratio, this can be enforced in a natural way by adding inequality constraints on the variables a, b, c such that:

$$\tilde{\mathbf{e}}_i = \arg\min_{\mathbf{e}} \|\tilde{\mathtt{M}}_i^c\mathbf{w}(\mathbf{e})\|_2^2 \quad s.t. \quad \begin{cases} a_l \leq a \leq a_u \\ b_l \leq b \leq b_u \\ c_l \leq c \leq c_u \end{cases} \qquad (23)$$

where $[a_l, b_l, c_l]$ and $[a_u, b_u, c_u]$ are respectively the lower and upper bounds on the three semi-axes a, b and c of the ellipsoid. Finally, we recover from $\tilde{\mathbf{e}}_i$ the estimated ellipsoid in the primal space, through Eqs. (21), (10), (18) and (11). The proposed regularisation takes into account the generic 3D proportion/scale properties of an object. In particular this can be helpful when having noisy 2D bounding box detections that could impact negatively on the 3D quadric estimation. As a consequence, this robustness could possibly lead to an improvement of the ellipse estimation in 2D if there are gross inaccuracies in the object detection.

5.1 Initialisation

Since the cost function in Eqs. (22) and (23) is non convex, a good initialisation is mandatory. As a first initialisation step, we computed the $\mathtt{SVD}$ solution from (17) in order to find the vector $\tilde{\mathbf{w}}_i^c$ which contains the elements defining the primal quadric $\tilde{\mathtt{Q}}_i^c$ and all the scale parameters $\tilde{\boldsymbol{\beta}}_i$. The

[3] Actually, the positivity of a^2, b^2, c^2 grants that $\mathtt{L}^*$ represents an ellipsoid and not a generic quadric.

main problem here is that the solution of the SVD represents a generic quadric surface, not in specific an ellipsoid. In light of this, we need to convert the generic quadric into an ellipsoid as a first attempt to obtain the initial value $\mathbf{e}_i^{(0)}$. This can be performed by extracting the axes using the following relation [33]:

$$\begin{bmatrix} a_i^{(0)} \\ b_i^{(0)} \\ c_i^{(0)} \end{bmatrix} = \sqrt{-\frac{\det \tilde{\mathtt{Q}}_i^c}{\det \tilde{\mathtt{Q}}_{i,33}^c} \begin{bmatrix} \lambda_1^{-1} \\ \lambda_2^{-1} \\ \lambda_3^{-1} \end{bmatrix}} \tag{24}$$

where $\tilde{\mathtt{Q}}_{i,33}^c$ is the 3×3 upper left submatrix of the primal quadric $\tilde{\mathtt{Q}}_i^c = adj^{-1}\left(\tilde{\mathtt{Q}}_i^{*c}\right)$, and λ_1, λ_2 and λ_3 are the eigenvalues of $\tilde{\mathtt{Q}}_{i,33}^c$.

Next, we can obtain the rotation matrix $\mathtt{R}(\boldsymbol{\theta}_i^{(0)})$, and consequently the Euler angles $\boldsymbol{\theta}_i^{(0)}$, considering that $\mathtt{R}(\boldsymbol{\theta}_i^{(0)})$ is equal to the matrix of eigenvectors of $\tilde{\mathtt{Q}}_{i,33}^c$. Finally, we obtain the translation vector $\mathbf{t}_i^{(0)}$ as follows:

$$\mathbf{t}_i^{(0)} = \tilde{\mathtt{Q}}_{i,33}^c \tilde{\mathtt{Q}}_{i,4}^c \tag{25}$$

where $\tilde{\mathtt{Q}}_{i,4}^c$ is a vector composed by the three upper right elements $[q_{14},\ q_{24},\ q_{34}]^\top$ of the matrix $\tilde{\mathtt{Q}}_i^c$. In general, one or more of the terms $a_i^{(0)}, b_i^{(0)}, c_i^{(0)}$ could be imaginary numbers, e.g. if the estimated quadric represents an hyperboloid. To overcome this problem, we use the modulus $|a_i^{(0)}|$, $|b_i^{(0)}|$ and $|c_i^{(0)}|$ as semi-axes length. This means to initialise the vector $\mathbf{e}_i^{(0)}$ as:

$$\mathbf{e}_i^{(0)} = [\boldsymbol{\theta}_i^{(0)}, \mathbf{t}_i^{(0)}, |a_i^{(0)}|, |b_i^{(0)}|, |c_i^{(0)}|, \tilde{\boldsymbol{\beta}}_i]. \tag{26}$$

5.2 Minimal cases: 2 and 3 views

In this section we discuss the minimal configuration under which the solution to the problem is unique (3 views) and the under constrained case in which some priors are needed to obtain a solution (two views).

In the ideal case of no errors, the equality constraint in Eq. (7) is fulfilled by a unique vector $\mathbf{w}_i$ (for less than a scale factor) only if the null space of the matrix $\mathtt{M}_i$ has dimension 1, that is, just one singular value of $\mathtt{M}_i$ is equal to zero. It can be demonstrated that this condition holds only if at least three views are present [28]. In case of three views the matrix $\mathtt{M}_i \in \mathbb{R}^{18\times 13}$ has rank 12 and hence just one singular value is equal to zero.

If the number of views decreases to two, the rank of matrix $\mathtt{M}_i \in \mathbb{R}^{12\times 12}$ is equal to 10. Thus, two singular values are equal to zero and the solution to (7) is given by a family of quadrics:

$$\hat{\mathtt{Q}}_i^*(\xi) = \xi\ \hat{\mathtt{Q}}_{1,i}^* + (1-\xi)\ \hat{\mathtt{Q}}_{2,i}^*, \tag{27}$$

where $\hat{\mathtt{Q}}_{1,i}^* \in \mathbb{R}^{4\times 4}$ and $\hat{\mathtt{Q}}_{2,i}^* \in \mathbb{R}^{4\times 4}$ are the dual quadrics obtained from the two right singular vectors of $\mathtt{M}_i$ associated to the singular values equal to zero, and ξ is a scalar parameter. Therefore two views are not sufficient to recover the correct solution, corresponding to a single ellipsoid, unless prior knowledge is injected in the problem.

To recover a reliable solution in this underconstrained case, we revert to the non-linear problem in Eq. (23). In such a way we can exclude quadrics different from ellipsoids and quadrics whose size or aspect ratio does not fulfil the inequality constraint on the axes, given by prior knowledge on the object shape and dimension.

A problem here arises on how to initialise (23) since in a real case every value of $\hat{\mathtt{Q}}_i^*(\xi)$, for an arbitrary ξ, could be in principle a good starting point. To solve for this issue, we adopted a greedy but feasible approach, evaluating $\hat{\mathtt{Q}}_i^*(\xi)$ over a number of values of ξ, obtained by discretising the interval $[0,1]$ into $K = 10$ equally spaced values[4]. For each selected $\hat{\mathtt{Q}}_i^*(\xi)$, we computed the corresponding initialisation vector $\mathbf{e}_i^{(0)}(\xi)$, according to the procedure described in Section 5.1. Finally, we solved the non-linear optimisation problem with boundary constraints on the ellipsoid axes, initialising it with $\mathbf{e}_i^{(0)}(\xi)$.

To this point, we have K potentially different ellipsoids $\tilde{\mathtt{Q}}_i(\xi)$, resulting from the K different initialisations. In order to choose the best among them, we selected the one maximising the intersection over union $O_{\mathtt{2D},i}(\xi)$ between the areas $\hat{\mathcal{C}}_{if}$ of the ellipses obtained from the bounding boxes and the areas $\tilde{\mathcal{C}}_{if}(\xi)$ of the ellipses reprojected from the estimated ellipsoid $\tilde{\mathtt{Q}}_i(\xi)$. In detail, defining:

$$O_{\mathtt{2D},i}(\xi) = \sum_{f=1}^{2} \frac{\hat{\mathcal{C}}_{if} \cap \tilde{\mathcal{C}}_{if}(\xi)}{\hat{\mathcal{C}}_{if} \cup \tilde{\mathcal{C}}_{if}(\xi)}, \tag{28}$$

we selected the ellipsoid $\tilde{\mathtt{Q}}_i(\bar{\xi}_i)$ with:

$$\bar{\xi}_i = \arg\max(O_{\mathtt{2D},i}(\xi)). \tag{29}$$

Despite ellipses from bounding boxes do not correspond exactly to ground truth reprojected ellipses, the metric $O_{\mathtt{2D}}$ was empirically found to have a good correlation with the best 3D ellipsoid reconstruction, in terms of volume overlap with the ground truth ellipsoid. On the contrary, we discarded the more trivial criterion of taking the solution for which the cost function in Eq. (23) was minimal.

6 Synthetic Experiments

We present extensive synthetic evaluation in order to asses the correctness of the proposed approach and to test its robustness to inaccuracies of the 2D bounding boxes position. In all the experiments, the accuracy of the estimated 3D localisation was measured by the volume overlap defined as:

$$O_{\mathtt{3D}} = \frac{1}{N} \sum_{i=1}^{N} \frac{\mathcal{Q}_i \cap \tilde{\mathcal{Q}}_i}{\mathcal{Q}_i \cup \tilde{\mathcal{Q}}_i}, \tag{30}$$

where $\mathcal{Q}_i$ and $\tilde{\mathcal{Q}}_i$ denote the volume of GT and estimated ellipsoids respectively.

When the i-th estimated quadric is not an ellipsoid, $O_{\mathtt{3D}}$ is set to zero. Notice that, when camera views are restricted to a small baseline or they are related to quasi-planar trajectories, $O_{\mathtt{3D}}$ could give poor results even with a small algebraic error in Eq. (9). However, we have chosen this metric since it measures in a direct way the success of the algorithm in recovering the 3D position and occupancy of an object. In the following we will denote the closed form solution obtained with SVD (Sec. 4.2, Eq. (17)) as LfD and the

[4] Since the quadric is defined up to a global scale factor, it is not necessary to consider values of ξ outside of the range $[0,1]$

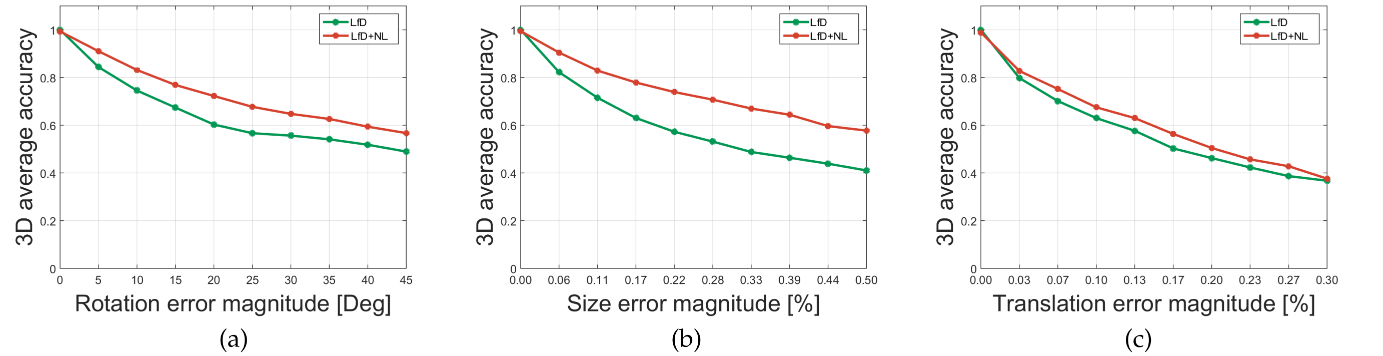


Fig. 4. Results for the synthetic tests versus different types of errors. Average accuracy given by LfD or LfD+NL for RE – rotation error (a), SE – size error (b), TE – translation error (c), measured by O_{3D} metric.

solution from non-linear optimisation (Sec. 5, Eqs. (22) and (23)) as LfD+NL. Where not specified, LfD+NL will indicate the non-linear solution without prior constraints in (22).

We tested our algorithms LfD and LfD+NL in a synthetic scenario with 50 ellipsoids randomly placed inside a cube of side 20 unit. The length of the largest axis L ranges from 3 to 12 units, according to a uniform PDF. The lengths of the other two axes are equal to γL with $\gamma \in [0.3, 1]$. Finally, axes orientation was fixed randomly. A set of 20 views were generated with a camera model that has a distance from the cube centre of 200 units and moves along a trajectory so that azimuth and elevation angles span the range $[0°, 60°]$ and $[0°, 70°]$ respectively[5]. Given the projection matrix $\mathtt{P}_f$ of each camera frame, GT ellipses given by the exact projections of the ellipsoids were calculated.

Synthetic tests were aimed at validating the robustness of the proposed method against common inaccuracies affecting object detectors, such as coarse estimation of the object centre, tightness of the bounding box with respect to the object size and variations over the object pose. Thus, each ellipse was corrupted by three errors, namely translation error (TE), rotation error (RE) and size error (SE), and fed to the proposed algorithm. To impose such errors, the ellipses centres coordinates t_1^c, t_2^c, the semi axes length l_1, l_2 and the orientation α of the first axis were perturbed as follows:

$$\hat{t}_j^c = t_j^c + \bar{l}\nu_j^t, \qquad \hat{\alpha} = \alpha + \nu^\alpha, \qquad \hat{l}_j = l_j\left(1 + \nu^l\right), \tag{31}$$

where ν_j^t, ν^α and ν^l are random variables with uniform PDF and mean value equal to zero, and $\bar{l} = (l_1 + l_2)/2$. In order to highlight the specific impact of each error, they were applied separately. Error magnitudes were set tuning the boundary values of the uniform PDFs of ν_j^t, ν^α and ν^l. In detail, for each kind of error, we considered 10 different values of ν_j^t, ν^α and ν^l, with uniform spacing, and we applied the resulting error realisations to the ellipses reprojections related to the 50 objects.

In Fig. 4 the average accuracy O_{3D} for both LfD and LfD+NL is displayed versus the error magnitude (i.e. the boundary value of the uniform PDF), for RE (Fig. 4(a)), SE (Fig. 4(b)) and TE (Fig. 4(c)). The results for both LfD and LfD+NL are perfect in absence of errors, as expected, and smoothly decrease as the three errors increase. In detail, O_{3D} for LfD+NL, drops from 1 (no errors) to 0.57, 0.59 and 0.38 for maximum RE, SE and TE respectively, while the accuracy of LfD drops from 1 to 0.49 (RE), 0.41 (SE) and 0.37 (TE) respectively. In general, LfD appears to be more sensitive to TE with respect to SE and RE.

The performance boost brought by non-linear optimisation is remarkable for SE, with an increase in O_{3D} of more than 0.2 over a wide range of SE errors (from 0.17 to 0.50). The performance boost is significant also for RE, yielding an increase in O_{3D} between 0.10 and 0.15 over the range of RE $[10, 45]$. Concerning TE, the improvement achieved by LfD+NL over LfD is more limited. However, one has to consider that robustness to translation errors is already diminished by data preconditioning, applied on both LfD and LfD+NL, thus making the contribution of non-linear optimisation less relevant. Notice also that the accuracy of 0.37, related to the maximum TE, is a reasonable value considering that an object detector placing the bounding box with a TE of 0.3 is typically judged to fail the detection[6].

Finally, it is to note that LfD+NL is generally more robust toward RE and SE (Figs. 4(a), (b)), than toward TE. The higher robustness toward RE and SE is quite important since such kind of errors are likely to happen very frequently whenever ellipses are fitted to bounding boxes. Even if the detector is accurate, the bounding box quantises the object alignment at steps of 90°, yielding a maximum RE of 45°. This tends to overestimate the object area, thus affecting SE, whenever the object is not aligned to the bounding box axes.

7 Real Experiments

We evaluate our approach in three standard datasets (ACCV, TUW and KITTI) for which we can obtain the ground truth location and occupancy of the objects in various scenarios. The imaging conditions, number of frames and camera paths are different for each dataset, thus evaluating the robustness of our approach under different realistic conditions. We also present a real experiment based on Google Tango where we use the odometry given by the device to obtain the pose of the calibrated camera. Given this

[5] Notice that the variation of object appearance due to such range of angle views can be handled by state-of-art object detectors [34]

[6] The TE is applied independently to both the horizontal and vertical components of the ellipse centre, i.e. see Eq. (31), resulting in a maximum overall translation of $0.3\sqrt{2}$.

information, our method locates objects in a scene while the user is moving. To further demonstrate the applicability of our approach, we have as well estimated performance when camera projection matrices are given by a self-calibration procedure.

Finally, to better evaluate our method, we have implemented two baseline approaches that find the 3D position and occupancy of the objects directly from the bounding boxes. The first baseline approach is the constraint propagation (CP) method developed by Farenzena et al. [35], which has been adapted to estimate, for each object, the polyhedrons given by the intersections of all the pyramids having the vertex on each camera centre and passing through the bounding boxes of the object detections. In this way, we can compare a cuboid based model against the proposed quadrics parametrisation for the object position and occupancy. The second baseline approach is a method based on Interval Analysis (IA) [36], [37], which solves a similar problem based on stereo triangulations. The main drawback of the IA is its tendency to overestimate the intersection volume. In fact, we have noticed that the error over the position of the 2D bounding box vertices amplifies greatly the extension of the 3D estimated interval.

Moreover, for the KITTI dataset we compare with competing 3D object localisation approaches on the dataset [38], [19][7].

Again, the closed form solution is named as LfD and the solution from non-linear optimisation as LfD+NL. Where not specified, LfD+NL will indicate the non-linear solution without prior constraints in (22). We also used the subscript c (LfD$_c$, LfD$_c$ + NL) to indicate the case when the camera parameters have been computed using self-calibration given by a hierarchical Structure from Motion pipeline [39].

7.1 ACCV dataset evaluation

The ACCV dataset [40] contains 15 sequences, each related to a single object laying on a table at different camera viewpoints (from 100 to 1000 per sequence). We selected the subset of 8 sequences for which the 3D point cloud of the object is provided, and limit the number of views to 100 for each sequence. For each object we evaluated the GT ellipsoid as the envelope of the 3D point cloud. Moreover, for each frame and each object, we generated a 2D bounding box by simulating the output of a multi-scale object detector providing bounding boxes with variable aspect ratio, like the well known Deformable Part Model (DPM) [41].

In Fig. 5, first row, an example of the localisation performance for the LfD method is displayed for the "Duck" sequence. The estimated ellipsoid almost perfectly fits the GT one in respect to location, size, eccentricity and alignment, as can be seen in three frames and in the overall 3D localisation. In Table 1 the accuracy for each sequence, for LfD, is reported in terms of $O_{\texttt{3D}}$.

Due to the large number of views and the good quality of the detections,the accuracy is on average quite good even without non-linear optimisation, with an $O_{\texttt{3D}}$ of 0.80. For this reason, non-linear optimisation did not improve significantly the results on this dataset with respect to the closed form solution. Using self-calibration, LfD$_c$ has a decrease in performance of a limited 12% on average. The image scenes in the dataset are well textured so it is possible to extract reliable feature points for the camera matrix estimation. The number of views also affects positively the CP baseline approach, as shown in the left image of Fig. 6: More the viewpoints span a large angle around the object, better the method performs. Even if these results are remarkable for CP, the average performance of LfD$_c$ and LfD is still higher in terms of $O_{\texttt{3D}}$. Finally, the IA approach overestimates the volume, as it can be seen in Fig. 6.

TABLE 1
$O_{\texttt{3D}}$ for sequences from ACCV dataset for LfD, LfD$_c$ and baseline.

	Iron	Duck	Ape	Can	Driller	Vise	Glue	Cat	**Avg**
LfD	**0.83**	**0.82**	**0.88**	**0.84**	0.66	**0.82**	**0.67**	**0.84**	**0.80**
LfD$_c$	0.67	0.72	0.85	0.70	0.49	0.68	0.61	0.70	0.68
CP	0.61	0.62	0.59	0.61	**0.69**	0.68	0.61	0.59	0.62
IA	0.29	0.12	0.27	0.31	0.31	0.49	0.36	0.12	0.28

7.2 TUW dataset evaluation

The TUW dataset [42] contains 15 annotated sequences showing a table with different sets of objects deployed on it. The number of frames per sequence ranges from 6 to 20. A 3D point cloud for each object is also provided. As for the ACCV dataset, we obtained the GT ellipsoids for each object and the 2D bounding boxes. We discarded sequences with strong occlusions that cannot be handled by current object detectors, and sequences where objects appear for a number of frames lower than 3, thus retaining 5 sequences. In this case, differently from the ACCV dataset, the objects are not centred in the 3D scene, then the initial 3D centring helps both the LfD and the LfD+NL to reach a better results in terms of localisation.

All the selected sequences have been tested with the methods LfD, LfD+NL, LfD$_c$, LfD$_c$+NL, CP and IA. The accuracy for each sequence is reported in Table 2, according to $O_{\texttt{3D}}$.

It can be noticed that the non-linear optimisation yields an improvement in the accuracy, in terms of $O_{\texttt{3D}}$, with respect to the closed form method on a great part of the tested sequences. In the case where the camera have been self-calibrated, there is a general decrease of performance of

TABLE 2
$O_{\texttt{3D}}$ for the sequences from TUW dataset.

	Seq.1	Seq.7	Seq.8	Seq.10	Seq11	**Avg**
LfD	0.43	**0.70**	0.74	0.77	**0.50**	0.63
LfD$_c$	0.34	0.40	0.34	0.62	0.28	0.40
LfD+NL	**0.49**	0.69	**0.75**	**0.79**	**0.50**	**0.64**
LfD$_c$ + NL	0.35	0.40	0.34	0.62	0.28	0.40
CP	0.10	0.16	0.18	0.22	0.14	0.16
IA	0.00	0.01	0.01	0.02	0.01	0.01

[7] For the ACCV and TUW datasets it was not possible to evaluate such algorithms because of the lack of enough training examples and the tested objects not being part of PASCAL VOC classes.

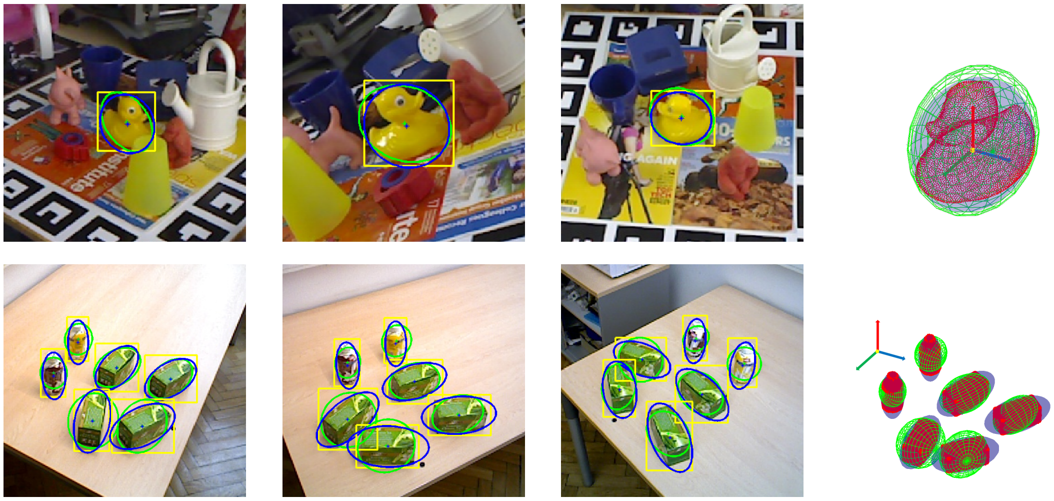


Fig. 5. Results for the ACCV (first row) and TUW (second row) datasets. The first three columns show a close up of the views with the output of a generic object detector (yellow bounding box) and projections of GT and estimated ellipsoids (blue and green ellipses respectively). The last column shows the cloud points of the object (red), GT ellipsoid (blue) and estimated ellipsoid (green).

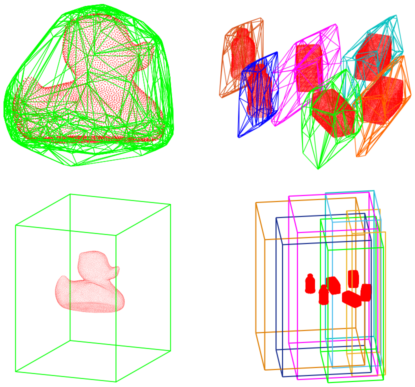


Fig. 6. Results for the ACCV and TUW sequences using the CP [35] (top images) and the IA [36] (bottom images) methods. In the figures we show the GT point clouds of the objects (red) and the estimated polyhedrons for the different objects. It can be noticed a tendency of both the approaches to overestimate the occupancy of the objects. This is more evident in the TUW dataset, where the IA fails the estimation.

about 20%, this possibly due to the presence of more homogeneous textures in this dataset (office environment). Notice also that there is not an appreciable difference between LfD$_c$ and LfD$_c$+NL. The results of the CP method are still below the performances of the LfD, LfD+NL, LfD$_c$ and LfD$_c$+NL due to the fewer number of camera poses, which clearly do not help to reduce the generated intersection volume, thus overestimating the occupancy of the objects, as it can be seen in the right image of Fig. 6. Clearly, the IA method fails at providing a solution, showing a high sensitivity to noise and few views of the object. Fig. 5, second row, shows an example of the localisation performance with the LfD + NL method for a selected sequence. The accuracy in the estimation of the objects' pose is remarkable and this trend is confirmed for all the other objects in the ACCV dataset in term of size, eccentricity and alignment of GT ellipsoids.

7.3 KITTI dataset evaluation and comparisons

The KITTI dataset [43] is composed by a set of sequences taken from a camera mounted on a moving car in an urban environment. The dataset provides full annotations for cars appearing in each frame from which GT ellipsoids can be computed. We sampled 5 sub-sequences displaying parked cars [8]. We generated 2D bounding boxes using the DPM object detector [41]. In particular, to detect the cars we used a Latent SVM model pre-trained by Geiger et al. [43] on the KITTI dataset.

In the KITTI dataset we only estimated the projection matrices P_f using [39], no ground truth information about the camera pose was available for this approach. Since we deal with an object category (the cars) for which priors on size and aspect ratio are well defined, we imposed boundary constraints on the ellipsoid axes, solving, for LfD$_c$+NL, the constrained problem as in Eq. (23) (see Section 5). In particular, we fixed for all the semi-axes a minimum value of 0.7 m and a maximum value of 3 m. We did not apply stricter constraints for each semi-axis in order to give the problem more freedom and decrease the probability of get stuck in local minima.

Ellipsoids estimation is particularly challenging on this dataset, since the camera motion is almost planar. Moreover, cars are usually placed at the street borders and the camera moves straight in most of the sequences. Hence, the range of angles between car and camera spanned by the sequence is very narrow and almost limited to the azimuth plane. Finally, each car appears in a limited subset of frames. This type of camera motion is very problematic for both CP and IA methods, since they are not able to constrain in depth the volume when the camera does not rotate around the objects.

In Table 3 quantitative results are displayed for the five selected sequences. Due to the difficulty of the dataset, LfD$_c$ achieves a limited performance in terms of $O_{\mathrm{3D}} = 0.05$;

[8] The selected sequences (Seq.) and the corresponding frames (Fr.) defining the sub-sequences are the following: Seq. 9 (Fr. 93 - 104); Seq. 22 (Fr. 49 - 85); Seq. 35 (Fr. 0 - 19); Seq. 36 (Fr. 43 - 63).; Seq. 39 (Fr. 129 - 159)

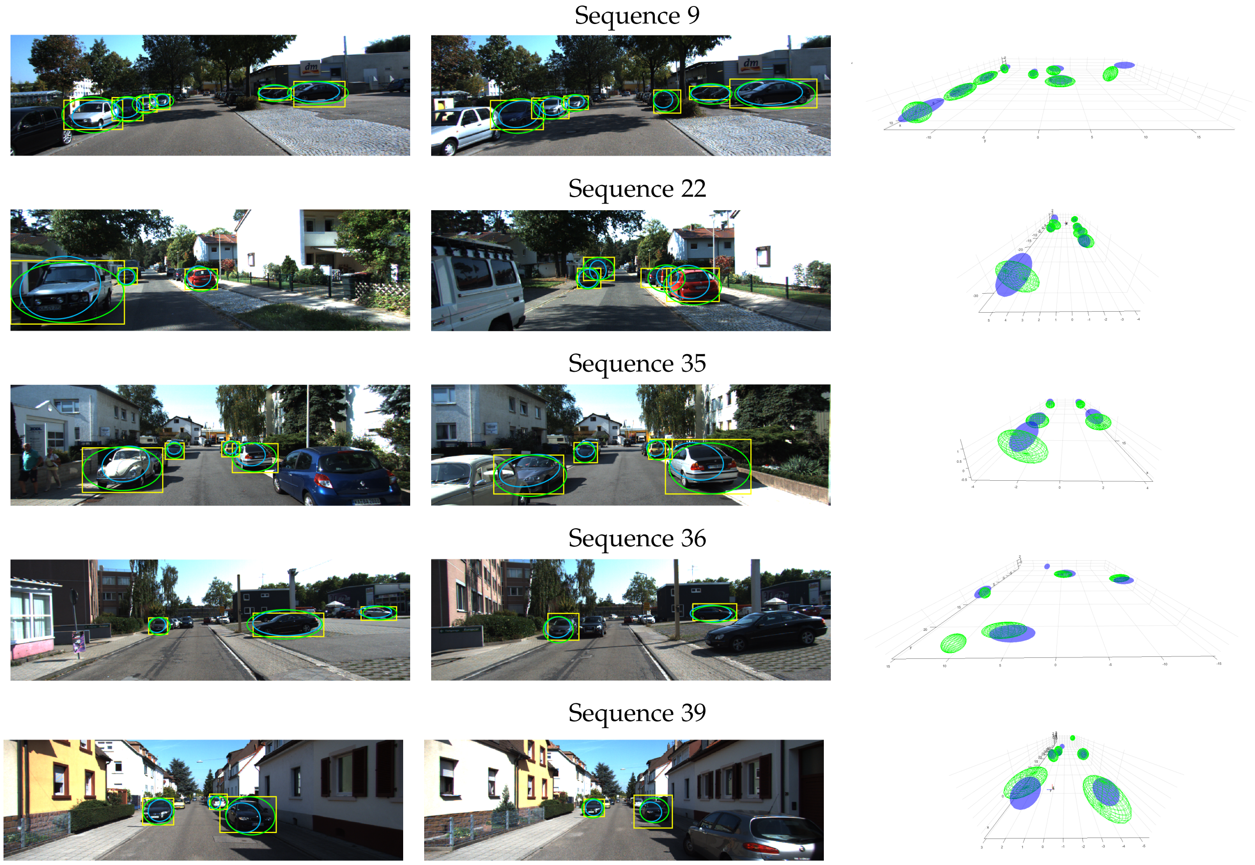


Fig. 7. Results for the five sequences of KITTI dataset. The left and centre images show a close up of the views with the output of a generic object detector (yellow bounding box) and projections of GT and estimated ellipsoids (blue and green ellipses respectively). The right images display the GT ellipsoids (blue) and estimated ellipsoids (green) in 3D.

however, a sharp improvement in accuracy is obtained with the non-linear optimisation LfD$_c$+NL, yielding an average $O_{3\text{D}} = 0.27$. The result for LfD$_c$+NL is visually confirmed by looking at Fig. 7, where, for each sequence, two representative frames and the 3D ground truth and estimated ellipsoids are displayed. Both CP and IA tend to grossly overestimate the volumes and they completely fail when estimating the occupancy.

TABLE 3
$O_{3\text{D}}$ for the sequences from the KITTI dataset.

	S.9	S.22	S.35	S.36	S.39	**Avg**
LfD$_c$	0.10	0.04	0.07	0.00	0.04	0.05
LfD$_c$ + NL	**0.30**	**0.32**	**0.15**	**0.23**	**0.36**	**0.27**
CP	0.00	0.00	0.00	0.00	0.00	0.00
IA	0.00	0.00	0.00	0.00	0.00	0.00

In detail, for Sequence 9 the estimated ellipsoids fairly fit the GT ones in six cases, while for the car on the top right and the bottom left the reconstruction has not high accuracy: this happens because in the first case the detections are very noisy, while in the second case there are few detections available (in the relative motion the car reaches the margin of the image, then goes out from the scene quickly). In Sequence 36 one of the cars is wrongly estimated due to the grossly inaccurate bounding boxes from detections, caused by the extreme distance of the car. In Sequence 39 there is just one case of failure on the localisation for the most distant car, for which very few and noisy detections are available.

Comparative Analysis We compared the proposed approach on the same sequences with two recent state of the art methods by Choi et al. [38] and Zia et al. [19] for 3D pose estimation. In detail, Choi et al. [38] propose a tracking algorithm able to estimate, given the bounding boxes provided by an object detector, the path of moving objects in space-time. The method employs an MCMC particle filter for the simultaneous estimation of the extrinsic camera parameters and of the position of the tracked elements in the 3D world frame.

Differently, the work of Zia et al. [19] can estimate jointly the shape and the pose of a specific object using priors on the context and on the object class. The priors on the context come from the monocular reconstruction of the ground plane, where the objects are supposed to lie, while the priors on the items are annotated manually from the dataset in the form of CAD models. The algorithm is properly tuned for the car class, in particular all the CAD models are generated from the KITTI dataset.

Since each method is fed by a different implementation of a DPM object detector (in [38] they used the release 3, while in [19] they trained their own model), in Table 4 we reported the number of detected objects for each sequence and each method. As can be seen the number of objects is nearly the same for our method (including both LfD_c and LfD_c+NL) and [38], and only slightly higher for [19], thus granting a meaningful comparison of the results.

In order to fairly compare the results of our method, we relied on a shared metric measuring the accuracy in the 3D object localisation. In particular we evaluated the percentage of estimated ellipsoid centres within 1 m and 2 m from the GT centroids of the objects. Results are reported in Table 5 for the four methods: LfD_c, LfD_c+NL, [38] and [19].

TABLE 4
Number of objects detected from each algorithm and each sequence.

	S.9	S.22	S.35	S.36	S.39
LfD_c	8	7	5	5	6
[38]	10	8	8	11	8
[19]	8	7	6	5	5

TABLE 5
Percentages of estimated centroids within 1 m or 2 m w.r.t. GT centroids for the 6 sequences of the KITTI dataset.

	S.9	S.22	S.35	S.36	S.39	**Avg**
LfD_c <1m	29	29	0	0	0	12
LfD_c+NL <1m	**86**	**86**	20	**40**	**67**	**60**
[38] <1m	1	13	2	1	1	4
[19] <1m	24	50	**53**	11	47	37
LfD_c <2m	57	43	20	20	17	31
LfD_c + NL <2m	**100**	**86**	60	**80**	**83**	**82**
[38] <2m	5	24	15	4	2	10
[19] <2m	57	65	**71**	24	72	58

LfD_c + NL achieved a notable performance with 82% of cars within 2 m and 60% within 1 m. Differently, LfD_c alone obtains lower results ($31\% < 2$ m and $12\% < 1$ m on average), again confirming the difficulty of the dataset and the importance of non-linear optimisation. Notice that, even for those quadrics estimated by LfD_c not corresponding to feasible ellipsoids, a quadric centre can be still computed. Results obtained with [38], are extremely low and, on average, worse than LfD_c and by far worse than LfD_c + NL for both the two metrics. Differently, results yielded by [19] are better than LfD_c, probably thanks to the strong priors given by the car CAD models, but significantly worse than LfD_c + NL, the latter outperforming all the other methods on average and on almost all the sequences.

In this dataset we also evaluated how precisely our algorithm can estimate the orientations of the cars by using the measure θ_{err}, which is the angle in radians between the main axes of estimated ellipsoids and GT ellipsoids. We compared the results given by LfD_c and LfD_c+NL with the orientations given by other three methods in Table 7.3. In particular, the comparing methods are a multi-view LSVM-based object detector trained by Geiger et al. [44], a multi-view fast soft cascade detection algorithm by Ohn-Bar et al. [45] and the method developed by Zia et al. [19]. All the comparing approaches, which can estimate the orientation through the appearance, are more accurate of 7.45 deg on average with respect to LfD_c and LfD_c + NL: In our geometrical methods the estimation of the orientation is strongly affected by the 2D errors on the bounding boxes.

TABLE 6
θ_{err} for the sequences from the KITTI dataset (in radiants).

	S.9	S.22	S.35	S.36	S.39	Avg
LfD_c	0.70	0.72	0.73	0.99	0.88	0.80
LfD_c + NL	0.52	0.44	0.54	0.41	0.26	0.43
[19]	0.46	**0.19**	**0.16**	0.35	0.39	0.31
[44]	**0.32**	0.31	0.25	0.52	0.20	0.32
[45]	0.43	0.34	0.26	**0.32**	**0.13**	**0.30**

Computation times. We compared the computation times of our two methods with the ones of [38] and [19], and with CP [35] and IA [36]. In detail, we evaluated the time required to process Sequence 9 of KITTI dataset running the MATLAB code on an Intel i7 with a 8-cores CPU running at 2.60 GHz. We can see from Table 7 that our two methods are comparable with [38], while are much faster with respect to [19]. We made explicit the timing T_c for the calibration step with the hierarchical Structure from Motion pipeline [39] in case of LfD, LfD+NL, CP and IA. The T_o is the optimisation time intrinsic to LfD and LfD+NL. LfD optimisation via `SVD` is performed in 0.60 s, while the LfD+NL lasted 12 s. Notice that [38] and CP are almost entirely optimised in C++, therefore a corresponding implementation of LfD+NL on C++ would highlight even more the computational advantage of our method. Finally, [19] has a probabilistic approach requiring high computation time: with respect to our methods it has between 1 and 2 orders of magnitude of difference.

TABLE 7
Computation time (in seconds) for each algorithm in case of the Sequence 9. The values T_c and T_o refer to the time spent for calibration [39] and optimisation with our method.

LfD	LfD + NL	[38]	[19]	CP	IA
24 T_c + 0.60 T_o	24 T_c + 12 T_o	34	2400	24 T_c + 1.33	24 T_c + 1.07

7.4 Minimal configurations test

In order to stress the method capabilities, we also tested minimal configurations of three and two views, the latter being solvable only with non-linear optimisation including inequality constraints, due to its ill-posed nature. We have chosen a set of views with a wide baseline (about 30 degrees in rotation) in order to make the tested case more challenging. Notice that in such condition standard methods based on feature point matching or disparity maps for 3D structure estimation are hardly applicable, due to the large variation in appearance that impairs the matching of points across different images.

Table 8 shows the results in terms of $O_{3\text{D}}$. In the case of 2 views, for all the objects of all the sequences, we found a $\bar{\xi}_i$ which maximises the intersection over union of the 2D reprojections and we applied the inequality constraints an the ellipsoid axes (as explained in the Sec. 5.2). The results confirm that in general accuracy is lower or slightly lower than in case of more views but still remaining reasonable for all the datasets. For the 3 views case and non-linear optimisation , the $O_{\mathtt{3D}}$ performance on average decreases of about 23% in the ACCV dataset and 7% in KITTI dataset with respect to non-minimal cases reported in the previous experiments, while it remains unchanged for TUW dataset. Moving to two views, the $O_{\mathtt{3D}}$ performance is further reduced of about 10%, 1% and 4% for ACCV, TUW and KITTI datasets respectively. For the ACCV dataset, the difference of $O_{\mathtt{3D}}$ in case of 3 views between the LfD and LfD+NL is not remarkable because, for each sequence, there is only one object at about the 3D coordinate centre. This makes the matrix $\tilde{\mathtt{Q}}_i^{*c}$ better conditioned, and even without the non-linear optimisation process a good reconstruction is achieved. For the other datasets the initial quadric centring is not always capable to find the correct centre location of each object. If the centring in the 3D coordinates is not accurate, the matrix $\tilde{\mathtt{Q}}_i^{*c}$ is not well conditioned as in the ACCV dataset, and the results with LfD are poor in terms of $O_{\mathtt{3D}}$; in general the results obtained by performing non-linear optimisation provide increments in performance coherent with the previous tests on non-minimal cases for the all the datasets.

In Table 9 we reported the percentage of cars for which the distance between the GT centroids and the estimated ones is < 1 m and < 2 m, for the KITTI dataset. In case of 3 views, the percentage of objects with a centroid < 1 m is 16% using LfD$_c$, and 51% using LfD$_c$+NL: In general, the non-linear optimisation improves the localisation of the objects if the initialisation is not far away from the exact solution. If we consider the case of 2 views, results are generally worse than with three views, for the LfD$_c$+NL solution. However the search over the values of parameter ξ (see Section 5.2) guarantees a good initialisation, limiting the number of outlying errors and leading consequently to better performance in terms of < 2 m on a couple of sequences (Seq. 35 and 39).

TABLE 8
Average $O_{\mathtt{3D}}$ in case of minimal configurations for the sequences from ACCV, TUW and KITTI dataset.

3 views

	ACCV	TUW	KITTI
LfD	0.66	0.53	0.07
LfD+NL	0.67	0.63	0.20

2 views

	ACCV	TUW	KITTI
LfD+NL	0.57	0.62	0.16

7.5 TANGO dataset

As a further evidence of the applicability of our method in real cases, we used the proposed LfD approach in combination with a tablet device equipped with a Google Tango system [46]. The device uses several sensor modalities (inertial measurement unit, a depth camera and a fisheye camera) in order to localise the tablet in indoor environments. This setup is a fitting test-bed for our method, since the device also outputs the corresponding projective camera matrix at each image frame.

TABLE 9
Percentages of estimated centroids within 1 m or 2 m w.r.t. GT centroids for the 5 sequences of the KITTI dataset in case of minimal configurations.

3 views

	S.9	S.22	S.35	S.36	S.39	**Avg**
LfD$_c$ <1m	13	29	**20**	**20**	0	16
LfD$_c$ <2m	25	57	**60**	40	33	43
LfD$_c$ + NL <1m	**63**	**71**	**20**	**20**	**83**	**51**
LfD$_c$ + NL <2m	**87**	**71**	**60**	**60**	**83**	**72**

2 views

	S.9	S.22	S.35	S.36	S.39	**Avg**
LfD$_c$+NL <1m	0	11	20	20	50	20
LfD$_c$ +NL <2m	38	67	80	20	100	61

To this end, we have implemented the LfD approach on the Tango tablet, in order to obtain a localisation of the objects while the user is navigating in an unknown area. Near real-time object detection was implemented on a server receiving timestamped images from the Tango device. The server continuously run a a Faster R-CNN [47] detector, this method represents an excellent compromise between detection accuracy and speed. Once the bounding boxes are extracted, they are matched in the video sequence using the tracking by detection method, developed by Geiger et al. [48] for the street road scenario of the KITTI dataset. For each sequence, a 3D point cloud is generated by the Tango system, which is used only for visualisation purposes and not for the localisation and object occupancy estimation. We considered the point cloud as a GT reconstruction of the scene, and we aligned the 3D CAD models of the objects to quantitatively evaluate the performances of LfD$_c$, CP and IA methods in terms of $O_{\mathtt{3D}}$.

Fig. 8(a) and 8(b) show respectively the tested indoor and outdoor sequences. The top images of Fig. 8 show two frames of each sequence, together with the 2D detections given by the faster R-CNN object detector and the reprojections of the estimated 3D quadrics. The bottom images of Fig. 8 display the estimated ellipsoids over the point cloud generated by the Tango device. As it can be noticed from the indoor sequence (Fig. 8(a)), the method can correctly localise the objects in the environment and the alignment with respect to the objects position is quite remarkable, also thanks to the camera estimates provided by the Tango. The average $O_{\mathtt{3D}}$ given by LfD$_c$ over all the objects is 0.32. In particular, the bottles and the toy car are well reconstructed, while the ellipsoid associated to the monitor has a larger depth than expected, due to the almost planar structure of the object that can not be estimated properly. The CP and IA methods do not provide any usable localisation because both of them are highly affected by errors over the detections. We have also tested the system in a challenging outdoor scenario, where Tango is supposed to provide worse camera poses since the device has been developed for being used indoor. From the object localisations in Fig. 8(b), it can be seen that the ellipsoids have a reasonable volumetric occupancy. The LfD$_c$ generates ellipsoids with a small displacement in depth over the position, possibly due to the drifting effect of the Tango system in the outdoor environment. Finally, LfD$_c$ obtains an O_{3D} of 0.19, which is similar to the O_{3D} of the CP. However, the CP can estimate the left car and the person, while it can not reconstruct the right car. As in the previous case, the IA fails completely the estimation.

TABLE 10
O_{3D} for the sequences from the TANGO dataset.

	Outdoor	Indoor
LfD$_c$	0.19	0.32
CP	0.18	0.00
IA	0.00	0.00

8 Discussions and Future Work

This paper presented a closed-form solution to recover the 3D occupancy of objects from 2D detections in multi-view. This algebraic solution is achieved through the estimation of a 3D quadric given 2D ellipsoids fitted at the objects detectors bounding boxes. Moreover a non-linear optimisation approach was devised to cope with possible ill-conditioning of the problem. The approach was tested against the common inaccuracies affecting object detectors such as coarse estimation of the object centre, tightness of the bounding box in respect to the object size and variations over the object pose. Experiments show that even with relevant errors, the estimated quadrics are able to localise the object in 3D and to define a reasonable occupancy.

Results on three different public datasets and two newly generated sequences using a Tango device demonstrate the practical applicability of the method and the ability to overcome baseline and state-of-art methods in terms of localisation accuracy when challenging conditions are present. The solutions of this problem has strong practical breakthroughs given the recent evolution of recognition algorithms. In particular, object detection is certainly going towards increased generality, so providing detectors for several object classes [49]. Thus, the proposed method can give a very efficient solution to leverage the 2D information for 3D scene understanding where objects can now be interrelated given their position in the metric space. This will inject important 3D reasoning in classic frameworks for object detection that are mostly restricted to 2D.

Regarding future work, mis-detections (i.e. outliers) might affect negatively the estimation of the quadrics. Thus, including further robustness in the optimisation by using additional information from trained multi-pose detector or exploiting the accuracy of a new generation of semantic segmentation [50] that might improve the performance of the system. This information would also greatly help in the estimation of the object orientation by avoiding the ambiguities of a purely geometrical approach.

Acknowledgements

The authors would like to thank the reviewers for their insightful comments and suggestions. We acknowledge Prof. Andrea Fusiello for making available code on Interval Analysis, Prof. Adrien Bartoli for helpful discussions on conics optimisation, Dr. Zeeshan Zia for running his object localisation code on our dataset and Dr. Pierre Fite-Georgel for valuable suggestions on Google Tango camera pose estimation.

References

[1] M. Ozuysal, V. Lepetit, and P. Fua, "Pose estimation for category specific multiview object localization," in *CVPR*. IEEE, 2009, pp. 778–785.

[2] D. Glasner, M. Galun, S. Alpert, R. Basri, and G. Shakhnarovich, "Viewpoint-aware object detection and pose estimation," in *ICCV*. IEEE, 2011, pp. 1275–1282.

[3] B. Pepik, P. Gehler, M. Stark, and B. Schiele, "3d2pm–3d deformable part models," in *ECCV*. Springer, 2012, pp. 356–370.

[4] J. Xiao, J. Chen, D.-Y. Yeung, and L. Quan, "Structuring visual words in 3d for arbitrary-view object localization," in *ECCV*. Springer, 2008, pp. 725–737.

[5] S. Savarese and L. Fei-Fei, "3d generic object categorization, localization and pose estimation," in *ICCV*. IEEE, 2007, pp. 1–8.

[6] S. Savarese and L. Fei Fei, "View synthesis for recognizing unseen poses of object classes," in *ECCV*. Springer, 2008, pp. 602–615.

[7] M. Torki and A. Elgammal, "Regression from local features for viewpoint and pose estimation," in *ICCV*. IEEE, 2011, pp. 2603–2610.

[8] F. Rothganger, S. Lazebnik, C. Schmid, and J. Ponce, "3d object modeling and recognition using local affine-invariant image descriptors and multi-view spatial constraints," *IJCV*, vol. 66, no. 3, pp. 231–259, 2006.

[9] E. Murphy-Chutorian and M. M. Trivedi, "Head pose estimation in computer vision: A survey," *PAMI*, vol. 31, no. 4, pp. 607–626, 2009.

[10] S. Vicente, J. Carreira, L. Agapito, and J. Batista, "Reconstructing pascal voc," in *CVPR*. IEEE, 2014, pp. 41–48.

[11] A. Kar, S. Tulsiani, J. Carreira, and J. Malik, "Category-specific object reconstruction from a single image," in *CVPR*. IEEE, 2015, pp. 1966–1974.

[12] D. Hoiem, A. A. Efros, and M. Hebert, "Putting objects in perspective," *IJCV*, vol. 80, no. 1, pp. 3–15, 2008.

[13] A. Gupta, A. A. Efros, and M. Hebert, "Blocks world revisited: Image understanding using qualitative geometry and mechanics," in *ECCV*. Springer, 2010, pp. 482–496.

[14] S. Y. Bao and S. Savarese, "Semantic structure from motion," in *CVPR*. IEEE, 2011, pp. 2025–2032.

[15] N. Fioraio and L. Di Stefano, "Joint detection, tracking and mapping by semantic bundle adjustment," in *CVPR*. IEEE, 2013, pp. 1538–1545.

[16] R. F. Salas-Moreno, R. A. Newcombe, H. Strasdat, P. H. Kelly, and A. J. Davison, "Slam++: Simultaneous localisation and mapping at the level of objects," in *CVPR*. IEEE, 2013, pp. 1352–1359.

[17] T. Kanade, "Recovery of the three-dimensional shape of an object from a single view," *Artificial intelligence*, vol. 17, no. 1-3, pp. 409–460, 1981.

[18] R. A. Brooks, "Model-based three-dimensional interpretations of two-dimensional images," *PAMI*, no. 2, pp. 140–150, 1983.

[19] M. Z. Zia, M. Stark, and K. Schindler, "Towards scene understanding with detailed 3d object representations," *IJCV*, vol. 112, no. 2, pp. 188–203, 2015.

[20] Z. Z. Muhammad, S. Michael, and S. Konrad, "Are cars just 3d boxes? jointly estimating the 3d shape of multiple objects," in *CVPR*. IEEE, 2014, pp. 3678–3685.

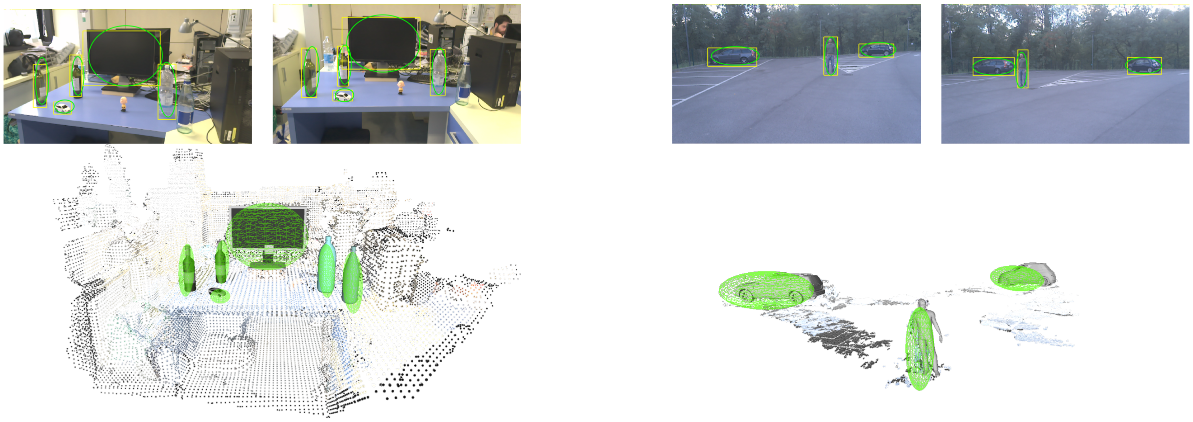
[21] M. Z. Zia, M. Stark, B. Schiele, and K. Schindler, "Detailed 3d representations for object recognition and modeling," *PAMI*, vol. 35, no. 11, pp. 2608–2623, 2013.

[22] J. Liebelt and C. Schmid, "Multi-view object class detection with a 3d geometric model," in *CVPR*. IEEE, 2010, pp. 1688–1695.

[23] S. Y. Bao, Y. Xiang, and S. Savarese, "Object co-detection," in *ECCV*. Springer, 2012, pp. 86–101.

[24] S. Y. Bao, M. Chandraker, Y. Lin, and S. Savarese, "Dense object reconstruction with semantic priors," in *CVPR*. IEEE, 2013, pp. 1264–1271.

[25] A. Dame, V. A. Prisacariu, C. Y. Ren, and I. Reid, "Dense reconstruction using 3d object shape priors," in *CVPR*. IEEE, 2013, pp. 1288–1295.



(a) (b)

Fig. 8. Results for the indoor (a) and outdoor (b) sequences of the TANGO dataset. In (a) and (b) the top images show two frames with the output of a generic object detector (yellow bounding box) and projections the estimated ellipsoids (green ellipses). The bottom images display the estimated ellipsoids (green), the 3D mesh vertexes provided by Tango and the registered 3D CADs of the objects.

[26] S. Fidler, S. Dickinson, and R. Urtasun, "3d object detection and viewpoint estimation with a deformable 3d cuboid model," in *NIPS*, 2012, pp. 611–619.

[27] S. Ma and L. Li, "Ellipsoid reconstruction from three perspective views," in *ICPR*, vol. 1. IEEE, 1996, pp. 344–348.

[28] G. Cross and A. Zisserman, "Quadric reconstruction from dual-space geometry," in *ICCV*. IEEE, 1998, pp. 25–31.

[29] M. Crocco, C. Rubino, and A. Del Bue, "Structure from motion with objects," in *CVPR*. IEEE, 2016, pp. 4141–4149.

[30] R. Hartley and A. Zisserman, *Multiple view geometry in computer vision*. Cambridge University Press, 2003.

[31] H. V. Henderson and S. Searle, "Vec and vech operators for matrices, with some uses in jacobians and multivariate statistics," *Canadian Journal of Statistics*, vol. 7, no. 1, pp. 65–81, 1979.

[32] T. Kato, *A short introduction to perturbation theory for linear operators*. Springer Science & Business Media, 2012.

[33] A. B. Ayoub, "The central conic sections revisited," *Mathematics Magazine*, pp. 322–325, 1993.

[34] S. K. Divvala, A. A. Efros, and M. Hebert, "How important are "deformable parts" in the deformable parts model?" in *ECCV-Workshops*. Springer, 2012, pp. 31–40.

[35] M. Farenzena and A. Fusiello, "3d surface models by geometric constraints propagation," in *CVPR*. IEEE, 2008, pp. 1–8.

[36] A. Fusiello, A. Benedetti, M. Farenzena, and A. Busti, "Globally convergent autocalibration using interval analysis," *PAMI*, vol. 26, no. 12, pp. 1633–1638, 2004.

[37] M. Farenzena, A. Fusiello, and A. Dovier, "Reconstruction with interval constraints propagation," in *CVPR*, vol. 1. IEEE, 2006, pp. 1185–1190.

[38] W. Choi and S. Savarese, "Multiple target tracking in world coordinate with single, minimally calibrated camera," in *ECCV*. Springer, 2010, pp. 553–567.

[39] M. Farenzena, A. Fusiello, and R. Gherardi, "Structure-and-motion pipeline on a hierarchical cluster tree," in *ICCV-Workshops*. IEEE, 2009, pp. 1489–1496.

[40] S. Hinterstoisser, V. Lepetit, S. Ilic, S. Holzer, G. Bradski, K. Konolige, and N. Navab, "Model based training, detection and pose estimation of texture-less 3d objects in heavily cluttered scenes," in *ACCV*. Springer, 2012, pp. 548–562.

[41] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan, "Object detection with discriminatively trained part-based models," *PAMI*, vol. 32, no. 9, pp. 1627–1645, 2010.

[42] A. Aldoma, T. Faulhammer, and M. Vincze, "Automation of ground truth annotation for multi-view rgb-d object instance recognition datasets," in *IROS*. IEEE, 2014, pp. 5016–5023.

[43] A. Geiger, P. Lenz, C. Stiller, and R. Urtasun, "Vision meets robotics: The kitti dataset," *IJRR*, vol. 32, no. 11, pp. 1231–1237, 2013.

[44] A. Geiger, C. Wojek, and R. Urtasun, "Joint 3d estimation of objects and scene layout," in *NIPS*, 2011, pp. 1467–1475.

[45] E. Ohn-Bar and M. M. Trivedi, "Fast and robust object detection using visual subcategories," in *CVPR-Workshops*. IEEE, 2014, pp. 179–184.

[46] "Tango project," https://get.google.com/tango/.

[47] S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," in *NIPS*, 2015, pp. 91–99.

[48] A. Geiger, M. Lauer, C. Wojek, C. Stiller, and R. Urtasun, "3d traffic scene understanding from movable platforms," *PAMI*, vol. 36, no. 5, pp. 1012–1025, 2014.

[49] T. Dean, M. A. Ruzon, M. Segal, J. Shlens, S. Vijayanarasimhan, and J. Yagnik, "Fast, accurate detection of 100,000 object classes on a single machine," in *CVPR*. IEEE, 2013, pp. 1814–1821.

[50] S. Zheng, S. Jayasumana, B. Romera-Paredes, V. Vineet, Z. Su, D. Du, C. Huang, and P. H. Torr, "Conditional random fields as recurrent neural networks," in *ICCV*. IEEE, 2015, pp. 1529–1537.

Cosimo Rubino received the BSc degree in Industrial Engineering and the MSc degree in Mechatronic Engineering in 2012 from University of Trento. He is currently PhD student in the Visual Geometry and Modelling (VGM) Lab at the PAVIS department of the Istituto Italiano di Tecnologia (IIT) in Genova. His research focuses in the areas of 3D scene understanding and motion segmentation.

Marco Crocco received the Laurea degree in electronic engineering in 2005 and the Ph.D. degree in electronic engineering, computer science and telecommunications in 2009 from the University of Genova. From 2005 to 2010, he was with the Department of Biophysical and Electronic Engineering (DIBE), University of Genova. From 2010 to 2015 he joined as a Post Doc the Pattern Analysis and Computer vision Department (PAVIS) and the Visual Geometry and Modeling (VGM) Lab at the Istituto Italiano di Tecnologia (IIT). His main research areas include array signal processing, machine learning and 3D scene understanding.

Alessio Del Bue received the Laurea degree in Telecommunication engineering in 2002 from University of Genova and his Ph.D. degree in Computer Science from Queen Mary University of London in 2006. He was a researcher in the Institute for Systems and Robotics (ISR) at the Instituto Superior Tecnico (IST) in Lisbon, Portugal. Currently, he is leading the Visual Geometry and Modelling (VGM) Lab at the PAVIS department of the Istituto Italiano di Tecnologia (IIT) in Genova. His research focuses in the areas of 3D scene understanding and non-rigid structure from motion.