3D Structure from Detections

Cosimo Rubino       Marco Crocco       Alessandro Perina       Vittorio Murino
Alessio Del Bue

Pattern Analysis and Computer Vision Department (PAVIS)
Istituto Italiano di Tecnologia
Via Morego 30, 16163 Genova, Italy

Abstract

We present a novel method to infer, in closed-form, the 3D spatial occupancy of a collection of rigid objects given 2D image detections from a sequence of images. In particular, starting from 2D ellipses fitted to bounding boxes, this novel multi-view problem can be reformulated as the estimation of a quadric (ellipsoid) in 3D. We show that an efficient solution exists in the dual-space using a minimum of three views. This algebraic solution can be negatively affected in the presence of gross inaccuracies in the bounding boxes estimation. To this end, we also propose a robust regularization method and a robust ellipse fitting algorithm able to improve performance in the presence of errors in the detected objects. Results on synthetic tests and on different real datasets, involving real challenging scenarios, demonstrate the applicability and potential of our method.

Localising objects in generic scenes is a fundamental step for higher level scene understanding. This task is of such importance that major efforts have been put in Computer Vision research in order to obtain efficient detectors from single images. These methods give remarkable results in finding several objects categories in natural scenes and thus putting the basis for the understanding of the visual world. However, while these results are certainly compelling, detection algorithms have been mostly restricted to the estimation of the objects position in 2D. When imaging the same scene from different viewpoints, a crucial question is how to deal with both 3D geometrical reasoning and higher-level object representation.

Recent works have clearly pointed out that this lack of 3D reasoning is limiting and 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. [13], the inclusion of 3D scene reasoning and rules has provided higher detection accuracies. Notably, attempts of unifying geometry and object representation have been achieved by defining elaborated Maximum a Posteriori (MAP) inference [2] or bundle adjustment with objects [7]. This way of pursuing high-level object reasoning in multi-view geometry has also inspired novel methods in Simultaneous Localisation and Mapping (SLAM) [16] where 3D objects are detected and localised in real time. Semantic information has been used, on the other hand, to infer the 3D shape of objects and the cameras viewpoints [20].

Differently, several works discarded multi-view relations attempting to solve 3D pose and shape estimation from a single image only. This severely under-constrained problem has to be solved with the use of strong semantic information (geometrical and physical constraints, i.e. object lying on the ground plane), 2D appearance and shape in the form of 3D wireframe or CAD models [17, 22, 23, 21]. Hejrati and Ramanan [10] instead use an analysis by synthesis approach that, guided by visual evidence, selects matching HOG patches which best represent a 3D object.

This work takes a different path from previous methods by showing that there is a sound and effective solution for
the 3D localisation of objects given only a set of image detections, i.e. 2D bounding boxes (BBs), using multi-view geometry constraints. To the best of the authors knowledge, this work is the first to answer this problem in 3D scene understanding. The solution proposed here recalls standard Structure from Motion (SfM) approaches [19, 18] where multiple views of the same rigid scene are used to obtain a 3D reconstruction given a set of 2D feature points measurements. Similarly we define a novel problem, namely Structure from Detection (SfD), that attempts to estimate an occupancy boundary that localise an object in 3D from a set of BBs obtained by general purpose detectors applied on multiple views (see Fig. [1] for a graphical representation). It is worth noting that the occupancy boundary encapsulates the estimation of object center, pose, size and aspect ratio in 3D. Moreover, differently from the large majority of works, we do not need geometric or semantic priors out of 2D bounding boxes from detections, nor advanced detectors yielding pose categorization.

In particular, we show that there exists an efficient closed-form solution to 3D object localisation if the problem is reformulated as the estimation of a quadric in 3D. This solution is not viable to derive using BBs (which is a piecewise-defined curve), but is mathematically feasible given a set of ellipses fitted to the original 2D BBs as extracted by the detectors. Given that the positions of the BBs can often be inaccurate, we propose a regularization method, based on priors on the object’s aspect ratios that fixes the possible ill-conditioning of the problem. Moreover we introduce a robust segmentation algorithm that can increase the reliability of the 2D ellipse fitting stage by finding the object profile inside the BBs. In such a way, improved results are achieved in challenging real scenarios on different freely available datasets.

The paper is structured as follows. Section 1 defines the problem and the related mathematical formalisation. Section 2 presents the SfD solution while Section 3 describes the regularization method and robust ellipse fitting approach. Experiments on real and synthetic data are discussed in Section 5 and then followed by concluding remarks in Section 6.

1. SfD problem statement

Let us consider a set of image frames $f = 1 \ldots F$ representing a 3D scene under different viewpoints. A set of $i = 1 \ldots N$ rigid objects is placed in arbitrary positions 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 BB $B_{if}$ given by a generic object detector. The BB is defined by a triplet of parameters: $B_{if} = \{w_{if}, h_{if}, c_{if}\}$, where $w_{if}$ and $h_{if}$ are two scalars defining the BB height and width respectively and $c_{if}$ is a 2-vector defining the BB center.

Our goal is to estimate the position and volume occupancy of each object in the 3D scene given the 2D BBs using multi-view constraints. In order to ease the mathematical formalization of the problem, we move from a BB representation of an object to an ellipsoid one. This is done by associating at each $B_{if}$ an ellipse fitting $\hat{C}_{if}$ that inscribes the BB, as shown in Fig. [2].

The aim of our problem is to find the 3D ellipsoids $Q_i$ whose projections on the image planes, associated to each frame $f = 1 \ldots F$, best fit the 2D ellipses $C_{if}$ in the image plane. 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:

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

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 $\hat{C}_{if} \in \mathbb{R}^{3 \times 3}$. The conic has five degrees of freedom, given by the six elements of the lower triangular part of the symmetric matrix $\hat{C}_{if}$ except one for the scale, since Eq. (1) is homogeneous in $\mathbf{u}$ [9]. Similarly to the ellipses, we represent the ellipsoids in the 3D space with the homogeneous quadratic form of a quadric equation:

$$\mathbf{x}^\top Q_i \mathbf{x} = 0,$$

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

Each quadric $Q_i$, when projected onto the image plane, gives a conic denoted with $C_{if} \in \mathbb{R}^{3 \times 3}$. The relationship between $Q_i$ and $C_{if}$ is defined by the projection matrices $P_f \in \mathbb{R}^{3 \times 4}$ associated to each frame. Such matrices can be estimated from the image sequence using standard self-calibration methods [9, 15]..

![Figure 2: Example of BBs (yellow) and corresponding fitted ellipses (red) for a set of objects.](Image 343x601 to 508x720)
2. Dual space fitting

Since the relationship between \( \mathbf{Q}_i \) and \( \mathbf{C}_{if} \) is not straightforward in the primal space, i.e. the Euclidean space of 3D points (2D points in the images), it is convenient to re-formulate it in dual space, i.e. the space of the planes (lines in the images) \( \mathbb{R}^n \). 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 \( \mathbf{Q}_i^* = \text{adj}(\mathbf{Q}_i) \), where \( \text{adj} \) is the adjoint operator, and the dual conic is defined by \( \mathbf{C}_{if}^* = \text{adj}(\mathbf{C}_{if}) \) \(^1\). Considering that the dual conic \( \mathbf{C}_{if}^* \), like the primal one, is defined up to an overall scale factor \( \beta_{if} \), the relation between a dual quadric and its dual conic projections \( \mathbf{C}_{if}^* \) can be written as:

\[
\beta_{if} \mathbf{C}_{if}^* = \mathbf{P}_f \mathbf{Q}_i^* \mathbf{P}_f^\top.
\]

In order to recover \( \mathbf{Q}_i^* \) in closed form from the set of dual conics \( \{ \mathbf{C}_{if}^* \}_{f=1}^{\cdots} \), we have to re-arrange Eq. (3) into a linear system. Let us define \( \mathbf{v}_i^* = \text{vech}(\mathbf{Q}_i^*) \) and \( \mathbf{c}_{if}^* = \text{vech}(\mathbf{C}_{if}^*) \) as the vectorization of symmetric matrices \( \mathbf{Q}_i^* \) and \( \mathbf{C}_{if}^* \) respectively. Then, let us arrange the products of the elements of \( \mathbf{P}_f \) and \( \mathbf{P}_f^\top \) in a unique matrix \( \mathbf{G}_f \in \mathbb{R}^{6 \times 10} \) as follows \(^2\):

\[
\mathbf{G}_f = \mathbf{D}(\mathbf{P} \otimes \mathbf{P})\mathbf{E}
\]

where \( \otimes \) is the Kronecker product and matrices \( \mathbf{D} \in \mathbb{R}^{6 \times 9} \) and \( \mathbf{E} \in \mathbb{R}^{16 \times 10} \) are two matrices such that \( \text{vech}(\mathbf{X}) = \mathbf{D} \text{vec}(\mathbf{X}) \) and \( \text{vec}(\mathbf{X}) = \mathbf{E} \text{vech}(\mathbf{X}) \) respectively, where \( \mathbf{X} \) is a symmetric matrix.

Given \( \mathbf{G}_f \), we can rewrite Eq. (3) as:

\[
\beta_{if} \mathbf{C}_{if}^* = \mathbf{G}_f \mathbf{v}_i^*.
\]

In order to get a unique solution for \( \mathbf{v}_i^* \) at least three image frames are needed. Therefore, stacking column-wise Eqs. (5) for \( f = 1 \ldots F \), with \( F \geq 3 \), we obtain the following linear system:

\[
\mathbf{M}_i \mathbf{w}_i = \mathbf{0}_b F
\]

where the matrix \( \mathbf{M}_i \in \mathbb{R}^{6F \times (10 + F)} \) and the vector \( \mathbf{w} \in \mathbb{R}^{10 + F} \) are defined as follows:

\[
\mathbf{M}_i = \begin{bmatrix}
G_1 & -c_{11} & 0_6 & 0_6 & \cdots & 0_6 \\
G_2 & 0_6 & -c_{22} & 0_6 & \cdots & 0_6 \\
G_3 & 0_6 & 0_6 & -c_{33} & \cdots & 0_6 \\
\vdots & 0_6 & 0_6 & 0_6 & \cdots & 0_6 \\
G_F & 0_6 & 0_6 & 0_6 & \cdots & -c_{iF}
\end{bmatrix}
\]

\[
\mathbf{w}_i = \begin{bmatrix}
\mathbf{v}_i^* \\
\beta_i
\end{bmatrix}
\]

with

\[
\beta_i = [\beta_{i1}, \beta_{i2}, \cdots, \beta_{iF}]^\top
\]

and \( \mathbf{0}_d \) being a zero column vector of length \( d \).

Note that in real cases the ellipses \( \mathbf{G}_i \) computed by a general purpose object detector might be inaccurate regarding the location of the BB and the window size. Likewise, this will have an effect on the ellipse fitting, inducing an error on the \( \mathbf{C}_{if} \). For this reason, if \( \mathbf{M}_i \) is the matrix given by real object detections we can find the solution by minimizing:

\[
\mathbf{w}_i = \arg \min_{\mathbf{w}_i} \| \mathbf{M}_i \mathbf{w}_i \|_2^2 \text{ s.t. } \| \mathbf{w}_i \|_2 = 1,
\]

were the equality constraint \( \| \mathbf{w}_i \|_2 = 1 \) avoids the trivial zero solution. The solution to the minimization problem in Eq. (9) can be obtained via SVD on the \( \mathbf{M}_i \) matrix, taking the right singular vector associated to the minimum singular value. The first 10 rows of \( \mathbf{w}_i \) are the vectorized elements of the estimated dual quadric denoted by \( \mathbf{v}_i^* \). To get back the estimated matrix of the quadric in the primal space, we obtain first the dual estimated quadric by \( \mathbf{\tilde{Q}}_i = \text{vech}^{-1}(\mathbf{v}_i^*) \), and subsequently apply the following relation:

\[
\mathbf{\tilde{Q}}_i = | \det \mathbf{\tilde{Q}}_i |^{\frac{1}{n}} (\mathbf{\tilde{Q}}_i)^{-1}.
\]

3. Regularisation and robust ellipse fitting

Image detectors can provide inaccurate results given occlusions, illumination variations and complex object poses. To this end, we propose a regularisation approach for Eq. (9) and a method to re-adjust the object orientation and size from the BBs extracted in the image sequence.

3.1. Regularized cost function

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 minimization is carried on the dual quadrics in Eq. (9) and the primal one is obtained by a matrix inversion as in Eq. (10). In the presence of noise the matrix \( \mathbf{Q}_i^* \) may become ill-conditioned and therefore small errors may cause relatively large errors in the estimated primal quadric. This problem is particularly evident when few camera views are spanning a limited range of angles. In such a case, the estimated quadric may result in a nearly degenerate ellipsoid, with dramatically unbalanced axes lengths, or even in wrong quadrics such as an hyperboloid.

To tackle this problem, we propose to add a regularization term that penalizes the departure of the estimated quadric from a sphere of a given size and center. In other words we enforce a prior on the aspect ratios of the objects modeled by the ellipsoids. A sphere in the dual space, centered at the origin, can be represented as a \( 4 \times 4 \) diagonal matrix with the first three diagonal elements positive and equal.
each other and the fourth diagonal element negative. To account for arbitrary translations, it is sufficient to pre- and post-multiply the diagonal matrix by a translation matrix. Hence a generic sphere in the dual space can be written as a function of a vector of five parameters $s = [t_1, t_2, t_3, a, b]^T$ as:

$$S^*(s) = TDT^T$$

with

$$T = \begin{bmatrix} 1 & 0 & 0 & t_1 \\ 0 & 1 & 0 & t_2 \\ 0 & 0 & 1 & t_3 \end{bmatrix}, \quad D = \begin{bmatrix} a & 0 & 0 & 0 \\ 0 & a & 0 & 0 \\ 0 & 0 & a & 0 \\ 0 & 0 & 0 & -b \end{bmatrix},$$

where $a > 0$ and $b > 0$. Thus, the regularization term can be defined as:

$$R(\hat{v}_i^*, s) = \|\hat{v}_i^* - vech(S^*(s))\|^2_2$$

(13)

where $\hat{v}_i^*$ is the vectorized $i$-th dual quadric normalized by minus its last (tenth) element giving:

$$\hat{v}_i^* = v_i^*/(-\hat{v}_i^*(10)).$$

(14)

Such normalization avoids the trivial solution when minimizing for Eq. (13). The minus sign has been set in order to preserve the coherence with the constraint $b > 0$. Finally, the solution to the regularized problem can be found solving for:

$$\hat{f} = \arg\min_f \|\hat{v}_i^* - \hat{w}_i\|^2_2 + \lambda R(\hat{v}_i^*, s) \quad s.t. \ a > 0, \ b > 0$$

(15)

where

$$f = \left[\hat{w}_i^\top \ s^\top\right]^\top, \quad \hat{w}_i = \left[\hat{v}_i^\top \ \beta_i^\top\right]^\top.$$  

(16)

Note that the normalization in Eq. (14) avoids the need for the quadratic equality constraint $\|\hat{w}_i\|^2_2 = 1$ in Eq. (15). Finally, the cost function in Eq. (15) is minimized with a nonlinear least squares procedure.

**Initialisation.** Since the convexity of the cost function cannot be guaranteed, a good initialization is mandatory, in particular for the elements of the dual quadric $\hat{v}_i^*$.

First of all we evaluated the sphere with the same center and the same volume of the quadric represented by $\hat{v}_i^*$. Next, we initialized both $\hat{v}_i^*$ and the parameters $s$ according to such a sphere. In detail, let us denote the starting guess values as $a^{(0)}$ and $b^{(0)}$ given by:

$$a^{(0)} = |e_{11} e_{22} e_{33}|^{-1/3}, \quad b^{(0)} = 1,$$

(17)

where $e_{11}, e_{22}$ and $e_{33}$ are the three elements of a vector $e_i$ defined as:

$$e_i = -\det(\tilde{Q}_i) e_i g(\tilde{Q}_i, 3, 3)$$

(18)

where $e_i g()$ is the operator that computes the eigenvalues of a square matrix and $\tilde{Q}_{i, 3 \times 3}$ is the $3 \times 3$ principal submatrix of the primal quadric $\tilde{Q}_i$ found by Eq. (10). If the initial $\tilde{Q}_i$ is not an ellipsoid, obviously the concept of volume preservation does not hold any more. However the initialization strategy, thanks to the modulus in Eq. (17) guarantees to start from a feasible solution corresponding to a sphere. The initialized translation terms $t_i^{(0)}, t_i^{(0)}$ and $t_3^{(0)}$ are set equal to the three translation parameters extracted by $\tilde{Q}_i$.

The initial vector $v_i^{(0)}$ is set equal to $vech(T^{(0)}D^{(0)}T^{(0)}\top)$, where $T^{(0)}$ and $D^{(0)}$ are defined by substituting $a^{(0)}, b^{(0)}, t_i^{(0)}, t_i^{(0)}$ and $t_3^{(0)}$ to the corresponding variables in Eq. (12). Finally $\beta_i^{(0)} = \beta_i / (-\hat{v}_i^*(10))$, where $\beta_i$ is the vector of scale factors corresponding to the elements of $\hat{w}_i$ from the 11-th to the last.

### 3.2. Ellipse fitting from image segmentation

In general, the BBs from generic object detectors are not precisely aligned with the true object center and often they include a relevant portion of background. More importantly, BB axes are aligned by construction to the image axes. Thus, the related ellipses are aligned to these axes as well: this results in a relevant rotation mismatch between the real ellipse enclosing the object in the image and the one given by the detector. This situation is even more complex when the ground truth ellipses have both strong axes rotation and eccentricity: the ground truth ellipses are badly approximated by the ellipse obtained from the BB, both in terms of rotation, shape and area (as in Fig. 3(a)).

![Figure 3](image-url)

(a) Example of mismatch between ground truth ellipse (blue) and ellipse from BB (red); (b) segmented object using SCA algorithm with ellipse fit to the segmentation (red) and ground truth ellipse (blue).

To cope with these limitations, we design an approach to obtain a tighter and more reliable ellipse fitting on the object inside the BB. In particular, we adopt a variation of the method proposed in [14], based on spatial correlations.

---

3 We recall that the volume of an ellipsoid is proportional to $\sqrt{1 / |e_{11} e_{22} e_{33}|}$ and the volume of a sphere is proportional to $\sqrt{a^3}$.  

---

4
among different image patches, to produce a binary segmentation mask. Here we exploit the fact that the pixel color and intensity distribution of the object in a set of frames are often sufficiently different from the background to allow an accurate object segmentation. Moreover as images come from a video sequence, the pose of the objects smoothly changes from one frame to another and these are the optimal conditions for stel component analysis - SCA algorithm [14]. SCA is a co-segmentation algorithm that segments images in S segments by learning K components of the model. Such components represent peculiar poses of an object and they are blended together to create a segmentation prior that well adapts to the different poses of the object in the image. The final segmentation is obtained by applying this flexible prior to a GMM-based segmentation.

Given the binary mask localising the object in each BB, we extract the object orientation and size from the covariance matrix Cov$_{if}$ of the 2D spatial distribution of the pixels belonging to the object $i$ in frame $f$. Now we can fit a 2D ellipse with axes aligned along the two eigenvectors of Cov$_{if}$ and with their magnitude proportional to the square root of its eigenvalues (i.e. $\lambda^j_{if}$ with $j = 1, 2$). In particular the magnitude of the ellipse’s axes is set making the second central moments of the region enclosed by the ellipse equal to the second central moments of the segmented region. An example of this segmentation process is displayed in Fig. 3(b) where the matching between ground truth and fitted ellipses is by far improved with respect to the ellipse from BB in Fig. 3(a).

4. Experiments

The proposed method has been tested on a synthetic scenario and on three real datasets. In all the tests, the accuracy of the results was measured by the volume overlap between ground truth and estimated ellipsoids ($O_{3D}$) defined as follows:

$$O_{3D} = \frac{1}{N} \sum_{i=1}^{N} \frac{Q_i \cap \tilde{Q}_i}{Q_i \cup \tilde{Q}_i}, \quad (19)$$

where $Q_i$ and $\tilde{Q}_i$ denote the volume of the ellipsoids given by the matrices $Q_i$ and $\tilde{Q}_i$, respectively. When the $i$-th estimated quadric is not an ellipsoid $\tilde{Q}_i$ is set to zero. Notice that, when camera views are restricted to a small range of angles or they are related to quasi-planar trajectories, $O_{3D}$ could give poor results even with a small algebraic error in Eq. (9). This might happen because of the ill-conditioning of the problem, yielding near degenerate solutions corresponding to ellipsoids with strong eccentricity. Nevertheless, we chose such 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 non-regularized solution obtained with SVD (Sec. 3.2) as SfD, the regularized solution (Sec. 3.1) as SfD+REG, the solution obtained from ellipses fitted to segmented images (Sec. 3.2) as SfD+ES (Ellipse Segmentation) and finally the solution implying both regularization and image segmentation as SfD+ES+REG.

4.1. Synthetic setup

The synthetic environment contains 50 objects modeled by ellipsoids randomly placed inside a cube of side 20 units. The length of the largest axis ranges from 3 to 12 units, according to a uniform Probability Density Function (PDF). The lengths of the other two axes are proportional to the largest axis with a proportionality factor ranging from 0.3 to 1 units. Finally, axes orientation was randomly generated.

A set of 20 views were generated with a camera distance from the cube center of 200 units and a trajectory was computed so that azimuth and elevation angles span the range $[0^\circ, 60^\circ]$ and $[0^\circ, 70^\circ]$ respectively. Given the projection matrix $P_f$ of each camera frame, ground truth 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 center, tightness of the BB 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 centers coordinates $c_1, c_2$, the axes length $l_1, l_2$ and the orientation $\alpha$ of the first axis were perturbed as follows:

$$\hat{c}_j = c_j + \bar{l}_\nu^j, \quad \hat{\alpha} = \alpha + \nu^\alpha, \quad \hat{l}_j = l_j \left(1 + \nu^l\right), \quad (20)$$

where $\nu^j, \nu^\alpha$ and $\nu^l$ 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 $\nu^j, \nu^\alpha$ and $\nu^l$. In Fig. 4 the average accuracy $O_{3D}$ of the proposed algorithm, for both SfD and SfD+REG, 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)). Concerning SfD+REG, the weighting parameter $\lambda$ has been kept equal for all the tests. The results for SfD are perfect in absence of errors but undergo a significant drop as the error increases. This confirms an expected sensitivity of the closed form solution to mismatches between fitted and ground truth ellipses. Nevertheless, when errors are moderate, the accuracy achieved is remarkable.

On the contrary, SfD+REG exhibits a very smooth decrease in accuracy versus the errors magnitudes, and a def-

---

1 Notice that the variation of object appearance due to such range of angle views can be handled by state-of-art object detectors.

2 We omit for simplicity the object and frame indexes.
initiately higher performance in comparison to SfD for almost all the error magnitudes, except for the case of zero error (RE, SE and TE) and the 5° RE. In detail, $O_{3D}$ for SfD+REG, drops from 0.52 (no errors) to 0.39, 0.33 and 0.16 for maximum RE, SE and TE respectively, while the accuracy of SfD drops from 1 to 0.15 (RE), 0.05 (SE) and 0.02 (TE) respectively.

Notice that both the methods are generally more robust toward RE and SE (Figs. 3(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 BBs. Even if the detector is accurate, the BB quantizes the object alignment at steps of 90°, yielding a maximum RE of 45°. This tends to over-estimate the object area, thus affecting SE, whenever the object is not aligned to the BB axes (see Fig. 3(a)). Concerning TE, we noticed that the robustness to such kind of errors decreases when ellipses are far from the image center and small in comparison to the image size. This is mainly due to the structure of the $3 \times 3$ matrix of the conic, whose entries become strongly unbalanced whenever the translation terms in the third row and column increase in respect to the $2 \times 2$ principal sub-matrix, leading likely to numerical problems. However, the maximum degree of failure for the regularized solution is reasonable considering that an object detector placing the BB with a TE of 0.3 is considered to fail the detection.

### 4.2. ACCV dataset evaluation

The ACCV dataset [12] 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 ground truth ellipsoid as the envelope of the 3D point cloud.

![Figure 4: Reconstruction performance for the synthetic tests versus different types of errors. Average accuracy given by SfD or SfD+REG for rotation error (a), size error (b), translation error (c), measured by $O_{3D}$ metric.](image)

For this reason, either adding regularization or fitting ellipses to segmented images did not improve significantly the results on this dataset.

| Object | SfD | SfD+REG |
|--------|-----|---------|
| Iron   | 0.71| 0.67    |
| Duck   | 0.83| 0.74    |
| Ape    | 0.47| 0.33    |
| Can    | 0.74| 0.67    |
| Driller| 0.34| 0.56    |
| Vise   | 0.67| 0.33    |
| Glue   | 0.33| 0.56    |
| Cat    | 0.56| 0.60    |

**Table 1:** $O_{3D}$ for sequences from ACCV dataset for SfD.

### 4.3. TUW dataset evaluation

The TUW dataset [1] has 15 annotated sequences showing a table with different sets of objects deployed over 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 ground truth ellipsoids for each object and the 2D BBs. 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 Fig. 6 an example of the reconstruction performance with the

---

6The TE is applied independently to both the horizontal and vertical components of the ellipse center, i.e. see Eq. (20), resulting in a maximum overall translation of $0.3\sqrt{2}$.
Figure 5: Duck reconstruction from ACCV [12] dataset, for SfD. (a), (b), (c): close up of three views with simulated output of a generic object detector (yellow BB) and projections of ground truth and reconstructed ellipsoids (blue and green ellipses respectively). (d): 3D shape of the object (red), ground truth ellipsoid (blue) and reconstructed ellipsoid (green).

SfD+ES+REG method for a selected sequence is displayed. The accuracy in the estimation of object location is remarkable for all the objects. The estimation of object size and shape, in term of size, eccentricity and alignment of ground truth ellipsoids, is also very accurate for the majority of objects. All the selected sequences have been tested with the four methods SfD, SfD+REG, SfD+ES and SfD+ES+REG. The accuracy for each sequence is reported in Table 2 according to $O_{3D}$. It can be noticed that the regularization yields a sharp improvement in the accuracy with respect to non regularized methods on all the tested sequences. In particular, the accuracy is even doubled moving from SfD to SfD+REG. The ellipse fitting from segmentation yields a further improvement, leading a 5% increment in accuracy moving from SfD+REG to SfD+ES+REG and a 3% increment moving from SfD to SfD+ES. Remarkably, the improvement from SfD+REG to SfD+ES+REG is achieved in every sequence.

Table 2: $O_{3D}$ for the sequences from TUW dataset.

| Seq. | SfD | SfD+ES | SfD+REG | SfD+REG+ES |
|------|-----|--------|---------|------------|
| 1    | 0.09| 0.12   | 0.48    | 0.50       |
| 2    | 0.43| 0.38   | 0.49    | 0.56       |
| 3    | 0.00| 0.16   | 0.48    | 0.52       |
| 4    | 0.60| 0.65   | 0.71    | 0.74       |
| 5    | 0.16| 0.69   | 0.40    | 0.43       |
| 6    | 0.25| 0.28   | 0.51    | 0.55       |
| Avg  | 0.25| 0.28   | 0.51    | 0.55       |

4.4. KITTI dataset evaluation

The KITTI dataset [8] 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 6 sub-sequences displaying parked cars. As for the previous datasets, we generated 2D BBs simulating a multi-scale object detector for each car and each frame. We pruned out those cars characterized by strong occlusions for which a reliable detection is unlikely. The total number of remaining cars was of 34 and the average number of views in which a car is visible is 10. 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. In such a condition the range of angles between car and camera spanned by the sequence of camera views is very narrow and almost limited to the azimuth plane. Finally, each car appears in a limited subset of frames. We did not apply segmentation on this dataset due to the extreme difficulty in segmenting some of the cars that are partially overlapping each other. In Table 3 quantitative results are displayed for the six selected sequences. Despite the difficulty of the dataset, SfD achieves a reasonable result of $O_{3D} = 0.17$ and the use of regularization almost doubles the accuracy, yielding $O_{3D} = 0.36$. The result for SfD+REG is visually confirmed by looking at Fig. 7: all the seven cars are correctly located in the 3D space, with a reasonable precision.

7The selected sequences (Seq.) and the corresponding frames (Fr.) defining the sub-sequences are the following: Seq 5 (Fr. 142 - 153); Seq 9 (Fr. 90 - 106); Seq. 22 (Fr. 49 - 86); Seq. 23 (Fr. 1 - 17); Seq. 35 (Fr. 1 - 5); Seq. 36 (Fr. 44 - 57).
Figure 7: Cars reconstruction from KITTI dataset for SfD+REG method. (a),(b): two views taken from sequence 9 with simulated output of a generic object detector (yellow BB) and projections of ground truth and reconstructed ellipsoids (blue and green ellipses respectively). (c): GT ellipsoid (blue) and reconstructed ellipsoid (green).

5. 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 BBs. Moreover a regularization 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 center, tightness of the BB in respect to the object size and variations over the object pose. To strengthen further the approach, a robust ellipse fitting method was introducing using a segmentation algorithm over the detected BBs in multi-view. Experiments show that even with relevant errors, the estimated quadrics are able to localise the object in 3D and to define a reasonable occupancy. Moreover, the proposed estimation of object orientation, by means of a segmentation algorithm, can be used in order to increase accuracy and the percentage of the reconstructed quadrics.

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. Thus, the proposed method can provide a quick and very efficient solution to leverage the 2D information for 3D scene understanding where objects can be inter-related 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 reasoning.

Regarding future work, mis-detections (i.e. outliers) might affect negatively the estimation of the quadrics. Thus, including further robustness in the optimization, through ad hoc regularization terms in the cost function, might improve the times within 2 m of the GT centroids, thus contributing to explain the very good performance of SfD.

Table 3: $O_{3D}$ for the sequences from the KITTI dataset.

|            | S.5 | S.9 | S.22 | S.23 | S.35 | S.36 | Avg. |
|------------|-----|-----|------|------|------|------|------|
| SfD        | 0.13| 0.05| 0.13 | 0.48 | 0.17 | 0.08 | 0.17 |
| SfD+REG    | 0.21| 0.49| 0.45 | 0.37 | 0.30 | 0.35 | 0.36 |

Table 4: Percentages of estimated centroids within 1 m or 2 m w.r.t. GT centroids for the 6 sequences of the KITTI dataset.

|          | S.5 | S.9 | S.22 | S.23 | S.35 | S.36 | Avg. |
|----------|-----|-----|------|------|------|------|------|
| SfD <1m   | 80  | 71  | 86   | 100  | 50   | 100  | 81   |
| SfD+REG <1m| 60 | 86  | 100  | 75   | 50   | 86   | 76   |
| SfD <2m   | 80  | 71  | 100  | 100  | 50   | 100  | 83   |
| SfD+REG <2m| 80 | 100 | 100  | 75   | 100  | 93   | 83   |
the performance of the system. Moreover, the ellipsoid orientations given the BBs can be further improved, especially when objects occlusions or similar appearance textures are present.

References

[1] A. Aldoma, T. Faulhammer, and M. Vincze. Automation of ground truth annotation for multi-view rgb-d object instance recognition datasets. In 2014 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), pages 5016–5023. IEEE, 2014.
[2] S. Y. Bao and S. Savarese. Semantic structure from motion. In 2011 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pages 2025–2032. IEEE, 2011.
[3] G. Cross and A. Zisserman. Quadric reconstruction from dual-space geometry. In Sixth International Conference on Computer Vision (ICCV), pages 25–31. IEEE, 1998.
[4] 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 2013 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pages 1814–1821. IEEE, 2013.
[5] S. K. Divvala, A. A. Efros, and M. Hebert. How important are "deformable parts" in the deformable parts model? In 2012 European Conference on Computer Vision (ECCV) . Workshops and Demonstrations, pages 31–40. Springer, 2012.
[6] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan. Object detection with discriminatively trained part-based models. IEEE Transactions on Pattern Analysis and Machine Intelligence, 32(9):1627–1645, 2010.
[7] N. Fioraio and L. Di Stefano. Joint detection, tracking and mapping by semantic bundle adjustment. In 2013 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pages 1538–1545. 2013.
[8] A. Geiger, P. Lenz, C. Stiller, and R. Urtasun. Vision meets robotics: The kitti dataset. International Journal of Robotics Research (IJRR), June 2013.
[9] R. Hartley and A. Zisserman. Multiple view geometry in computer vision. Cambridge University Press, 2003.
[10] M. Hejrati and D. Ramanan. Analysis by synthesis: 3d object recognition by object reconstruction. In 2014 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pages 2449–2456. IEEE, 2014.
[11] H. V. Henderson and S. Searle. Vec and vech operators for matrices, with some uses in jacobians and multivariate statistics. Canadian Journal of Statistics, 7(1):65–81, 1979.
[12] 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 2012 Asian Conference on Computer Vision (ACCV), pages 548–562. Springer, 2012.
[13] D. Hoiem, A. A. Efros, and M. Hebert. Putting objects in perspective. International Journal of Computer Vision, 80(1):3–15, 2008.
[14] N. Jojic, A. Perina, M. Cristani, V. Murino, and B. Frey. Stel component analysis: Modeling spatial correlations in image class structure. In 2009 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pages 2044–2051. IEEE, 2009.
[15] M. Pollefeys, R. Koch, and L. Van Gool. Self-calibration and metric reconstruction inspite of varying and unknown intrinsic camera parameters. International Journal of Computer Vision, 32(1):7–25, 1999.
[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 2013 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pages 1352–1359. IEEE, 2013.
[17] N. Silberman, D. Hoiem, P. Kohli, and R. Fergus. Indoor segmentation and support inference from rgbd images. In 2012 European Conference on Computer Vision (ECCV), pages 746–760. Springer, 2012.
[18] P. Sturm and B. Triggs. A factorization based algorithm for multi-image projective structure and motion. In 1996 European Conference on Computer Vision (ECCV), pages 709–720. Springer, 1996.
[19] C. Tomasi and T. Kanade. Shape and motion from image streams under orthography: a factorization method. International Journal of Computer Vision, 9(2):137–154, 1992.
[20] S. Vicente, J. Carreira, L. Agapito, and J. Batista. Reconstructing pascal voc. In 2014 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pages 41–48. IEEE, 2014.
[21] Y. Xiang and S. Savarese. Estimating the aspect layout of object categories. In 2012 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pages 3410–3417. IEEE, 2012.
[22] M. Z. Zia, M. Stark, B. Schiele, and K. Schindler. Detailed 3d representations for object recognition and modeling. IEEE Transactions on Pattern Analysis and Machine Intelligence, 35(11):2608–2623, 2013.
[23] M. Z. Zia, M. Stark, and K. Schindler. Are cars just 3d boxes? jointly estimating the 3d shape of multiple objects. In 2014 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pages 3678–3685, 2014.