QuadricSLAM: Dual Quadrics from Object Detections

Lachlan Nicholson, Michael Milford, and Niko Sünderhauf

***Abstract*— Research in Simultaneous Localization And Mapping (SLAM) is increasingly moving towards richer world representations involving objects and high level features that enable a semantic model of the world for robots. Many of these advances are grounded in state-of-the-art computer vision techniques primarily developed in the context of image-based benchmark datasets, leaving several challenges to be addressed in adapting them for use in robotics. In this work, we derive a SLAM formulation that uses dual quadrics as 3D landmark representations, exploiting their ability to efficiently represent the size, position and orientation of an object, and show how 2D bounding boxes (such as those typically obtained from visual object detection systems) can directly constrain the quadric parameters via a novel geometric error formulation. We develop a sensor model for deep-learned object detectors that addresses the challenge of partial object detections often encountered in robotics applications, and demonstrate how to jointly estimate the camera pose and constrained dual quadric parameters in factor graph based SLAM.**

I. Introduction

In recent years, impressive vision-based object detection performance improvements have resulted from the "rebirth" of Convolutional Neural Networks (ConvNets). Despite these impressive developments, the Simultaneous Localization And Mapping community (SLAM) has not yet fully adopted the newly arisen opportunities to create semantically meaningful maps. SLAM maps typically represent *geometric* information, but do not carry immediate object-level *semantic* information. Semantically-enriched SLAM systems are appealing because they increase the richness with which a robot can understand the world around it, and consequently the range and sophistication of interactions that robot may have with the world, a critical requirement for their eventual widespread deployment at workplaces and in homes.

Semantically meaningful maps should be object-oriented, with objects as the central entities of the map. *Quadrics*, i.e. 3D surfaces such as ellipsoids, are ideal landmark representations for object-oriented semantic maps. Quadrics have a very compact representation, can be manipulated efficiently within projective geometry, and capture information about the size, position, and orientation of an object.

The link between object detections and dual quadrics was recently investigated by [1], [2] and [3]. However, previous work utilized quadrics as a parametrization for landmark *mapping* only [2], was limited to an orthographic camera [1], or used an *algebraic* error that proved to be invalid when landmarks are only partially visible [3]. In this work we formulate a novel geometric error that is well-defined even when the observed object is only partially visible in the image. Furthermore, we investigates the utility of quadric based landmarks in a factor graph SLAM formulation that jointly estimates camera poses and quadric parameters from noisy odometry and object detection bounding boxes using a general perspective camera.

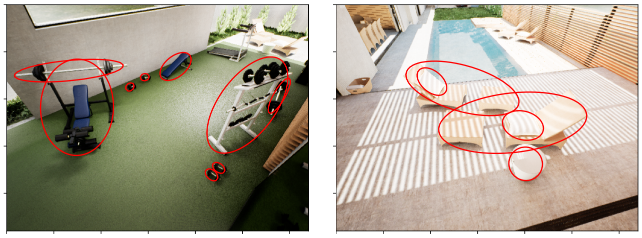

Fig. 1: QuadricSLAM uses *objects* as landmarks and represents them as constrained dual quadrics in 3D space. This figure depicts the estimated quadrics fit to true objects, with red ellipses as the 2D outline of the 3D quadric surfaces.

II. Dual Quadrics – Fundamental Concepts

Quadrics are surfaces in 3D space that are defined by a 4×4 symmetric matrix $\mathbf{Q}$, so that all points $\mathbf{x}$ on the quadric fulfill $\mathbf{x}^\mathsf{T}\mathbf{Q}\mathbf{x}=0$. Examples for quadrics are bodies such as spheres, ellipsoids, hyperboloids, cones, or cylinders.

When a quadric is projected onto an image plane, it creates a dual *conic*, following the simple rule $\mathbf{C}^* = \mathbf{P}\mathbf{Q}^*\mathbf{P}^\mathsf{T}$. Here, $\mathbf{P} = \mathbf{K}[\mathbf{R}|\mathbf{t}]$ is the camera projection matrix that contains intrinsic and extrinsic camera parameters. Conics are the 2D counterparts of quadrics and form shapes such as circles, ellipses, parabolas, or hyperbolas.

III. A Sensor Model for Modern Object Detectors

A. Motivation

Our goal is to incorporate state-of-the-art deep-learned object detectors such as [4]–[6] as a *sensor* into SLAM. We thus have to formulate a *sensor model* that can predict the observations of the object detector given the estimated camera pose $\mathbf{x}_i$ and the estimated quadric parameters $\mathbf{q}_j$.

We therefore seek a formulation for the sensor model $\boldsymbol{\beta}(\mathbf{x}_i,\mathbf{q}_j)=\hat{\mathbf{d}}_{ij}$, mapping from camera pose $\mathbf{x}_i$ and quadric $\mathbf{q}_j$ to predicted bounding box observation $\hat{\mathbf{d}}_{ij}$. This sensor model allows us to formulate a *geometric error* term between the predicted and observed object detections.

The authors gratefully thank John Skinner for his contributions to the evaluation environment. This research was conducted by the Australian Research Council Centre of Excellence for Robotic Vision (project number CE140100016). Michael Milford is supported by an Australian Research Council Future Fellowship (FT140101229). The authors are with the ARC Centre of Excellence for Robotic Vision, Queensland University of Technology (QUT), Brisbane, Australia.

Contact: `lachlan.nicholson@hdr.qut.edu.au`

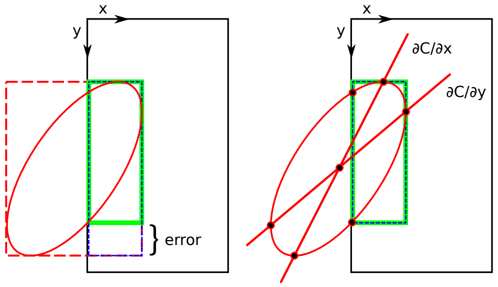

Fig. 2: Visual comparison between an incorrect bounding box prediction (left) and a correct sensor model (right), here the red ellipse represents a partially visible object and the expected bounding box from an object detector is highlighted in green.

B. Deriving the Object Detection Sensor Model $\boldsymbol{\beta}$

Our derivation of $\boldsymbol{\beta}(\mathbf{x}_i, \mathbf{q}_j) = \hat{\mathbf{d}}_{ij}$ starts with projecting the estimated quadric parametrized by $\mathbf{q}_j$ into the image using the camera pose $\mathbf{x}_i$ according to $\mathbf{C}^*_{ij} = \mathbf{P}_i \mathbf{Q}^*_{(\mathbf{q}_j)} \mathbf{P}_i^\mathsf{T}$ with $\mathbf{P} = \mathbf{K}[\mathbf{R}|\mathbf{t}]$ comprising the intrinsic ($\mathbf{K}$) and pose parameters of the camera. Given the dual conic $\mathbf{C}^*$, we obtain its primal form $\mathbf{C}$ by taking the adjugate.

A naive sensor model would simply calculate the enclosing bounding box of the conic $\mathbf{C}$ and truncate this box to fit the image. However, as illustrated in Figure 2, this can introduce significant errors when the conic's extrema lie outside of the image boundaries.

An accurate sensor model requires knowledge of the intersection points between conic and image borders. The correct prediction of the object detector's bounding box therefore is the minimal axis aligned rectangle that envelopes all of the conic contained within the image dimensions. We will explain the correct method of calculating this conic bounding box, denoted $\texttt{BBox}(\mathbf{C})$, below. The overall sensor model is then defined as

$$\boldsymbol{\beta}(\mathbf{x}_i, \mathbf{q}_j) = \texttt{BBox}\left(\mathrm{adjugate}(\mathbf{P}\mathbf{Q}^*_{(\mathbf{q}_j)}\mathbf{P}^\mathsf{T})\right) = \hat{\mathbf{d}}_{ij} \qquad (1)$$

C. Calculating the On-Image Conic Bounding Box

We can calculate the correct on-image conic bounding box by the following algorithm which we denote $\texttt{BBox}(\mathbf{C})$:

1) Find the four extrema points of the conic $\mathbf{C}$, i.e. the points $\{\mathbf{p}_1, ..., \mathbf{p}_4\}$ on the conic that maximise or minimise the x or y component respectively.
2) Find the up to 8 points $\{\mathbf{p}_5, ..., \mathbf{p}_{12}\}$ where the conic intersects the image boundaries.
3) Remove all non-real points and all points outside the image boundaries from the set $\mathcal{P} = \{\mathbf{p}_1, ..., \mathbf{p}_{12}\}$.
4) Find and return the maximum and minimum x and y coordinate components among the remaining points.

IV. Experiments and Evaluation

We implemented the SLAM problem as a factor graph where the robot poses and dual quadrics, X^* and Q^*, populated the latent variables of the graph, connected with odometry factors U and 2D bounding box factors D. Given a set of noisy odometry measurements and noisy bounding box observations, we estimate the optimal camera trajectory and object landmark parameters by minimizing the odometry error ($\|f(\mathbf{x}_i, \mathbf{u}_i) \ominus \mathbf{x}_{i+1}\|^2_{\Sigma_i}$) and the bounding box error ($\|\mathbf{d}_{ij} - \boldsymbol{\beta}_{(\mathbf{x}_i,\mathbf{q}_j)}\|^2_{\Lambda_{ij}}$). We evaluate the resulting trajectory and landmark parameters in a simulation environment of 250 trajectories, comparing the odometry estimate, initial quadric solution, and SLAM solution to the ground truth camera trajectory and each objects 3D axis-aligned bounding boxes.

TABLE I: Comparison of the average RMSE errors for the trajectory and landmark position (cm), as well as the landmark shape and quality defined by the centered Jaccard distance and the standard Jaccard distance respectively.

	ATE$_{\text{trans}}$	LM$_{\text{trans}}$	LM$_{\text{shape}}$	LM$_{\text{quality}}$
Odometry	58.95	-	-	-
SVD solution	-	57.86	0.61	0.85
QuadricSLAM	**20.49**	**17.14**	**0.44**	**0.59**

V. Results and Conclusions

We summarize the results of our experiments in Table I. The results show that quadric landmarks significantly improve the quality of the robot trajectory and the estimated map, providing accurate high level information about the shape and position of objects within the environment. Explicitly, QuadricSLAM gains a 65.2% improvement over the initial trajectory estimate and a 70.4%, 26.7% and 30.6% improvement on initial landmark positions, shape and quality.

These improvements provide justification to the use of quadric landmarks as coarse object representations, and a first step towards object-oriented semantic SLAM. Using noisy 2D bounding boxes such as those typically provided by standard object detectors, we are able to constrain the parameters of dual quadric landmarks, reobserving these landmarks reduces the effect of odometry drift.

Finally, we address the issue of partial object detections by defining a sensor model for modern object detectors that aims to predict the bounding box we would expect to see given the current pose and map estimates.

References

[1] M. Crocco, C. Rubino, and A. Del Bue, "Structure from motion with objects," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 4141–4149.

[2] C. Rubino, M. Crocco, and A. Del Bue, "3d object localisation from multi-view image detections," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2017.

[3] N. Sünderhauf and M. Milford, "Dual quadrics from object detection boundingboxes as landmark representations in slam," *arXiv preprint arXiv:1708.00965*, 2017.

[4] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," in *Advances in Neural Information Processing Systems (NIPS)*, 2015, pp. 91–99.

[5] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A. C. Berg, "SSD: Single shot multibox detector," in *European conference on computer vision*. Springer, 2016, pp. 21–37.

[6] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask r-cnn," *arXiv preprint arXiv:1703.06870*, 2017.