Multi-camera VSLAM: from former information losses to self-calibration

Joan Solà

Abstract— **Visual SLAM is, in recent years, a very active research area, the result of activities in the convergence of the robotics and computer vision communities. We present an overview of techniques, from classical filtering to bundle adjustment solutions, for both monocular and stereo (or multicamera) systems, and emphasize that classical SLAM solutions have been discarding precious sensory information. In particular, the ability of vision to sense objects at infinity should be exploited at its maximum because it is precisely these remote objects that will provide long-term, stable angular references (in the way a compass would do). Monocular SLAM systems have already solved this issue, but stereo and multicamera systems have not. We propose in these cases to use monocular SLAM algorithms using fusion to incorporate all the information. Numerous advantages like desynchronization of the sensors firing, the possibility of using several unequal cameras, or selfcalibration, will naturally arise. We develop a particular method for extrinsically decalibrated stereo systems to illustrate the proposed ideas. We evaluate the method with a real indoor experiment, and highlight and discuss both its assets and drawbacks.**

I. Introduction

The Simultaneous Localization And Mapping (SLAM) problem, as formulated by the robotics community, is that of creating a *map* of the perceived environment while getting *localized* in it, two tasks that are integrated in such a way as to benefit each other: good localization is crucial to creating good maps, and a good map is crucial to becoming localized. That is why they must be performed *simultaneously*, and hence the full acronym SLAM. In recent years the maturity of both online algorithms for SLAM, and fast and reliable image processing tools to extract the relevant geometrical information out of images, coming from the computer vision literature, have crystallized into a considerable number of real-time demonstrations of visual SLAM (VSLAM).

This work relates mostly to the estimation side of SLAM when dealing with vision. The exclusive properties of visual sensing (bearings-only, infinity range and rich appearance information) have their impact on the proper estimation procedures to use. We have mainly three objectives:

- Provide an informal yet coherent panorama of the different estimation methodologies suitable to be used in VSLAM, highlighting the key aspects to show that the most popular VSLAM systems – *i.e.*, those using filtering – have been historically discarding precious sensory information. In particular, a proper visual SLAM system should be able to exploit information a) from the first observation, b) independently of the vehicle trajectory, and c) up to the infinity range.
- Show that it is possible to exploit this information to extend VSLAM capabilities beyond localization and mapping. In particular, show that we have enough observed degrees of freedom to be used for sensor self-calibration. This improves the sensors' precision and therefore the overall map and localization accuracies.
- Convey the idea that the well-known concept of sensor fusion, when applied to multicamera SLAM, provides the necessary framework to achieve all these assets. This is demonstrated with a simple stereo-SLAM solution with self-calibration based on the extended Kalman filter (EKF).

This text is organized as follows: Section II is a brief overview of VSLAM techniques. Section III revises the necessary material for monocular SLAM and presents the main ideas that will be exploited later. Section IV explains how to set up Bi-Camera SLAM, an application for stereo benches with self-calibration. Section V presents experimental results and Section VI presents conclusions and future directions.

II. VSLAM methods and related techniques

A. Monocular SLAM and Structure From Motion (SFM)

Possibly the best example of the aforementioned technological crystallization is monocular SLAM. In monocular SLAM, good 3D landmark estimates require several observations from different vantage points. This makes landmark initialization difficult, to the point that satisfactory methods, able to exploit the whole geometrical information provided by the cameras, have been possible only recently. Historically, *delayed landmark initialization* [1], [2] methods tried to obtain a full 3D estimate before initialization via several observations. Delayed schemes can only initialize landmarks with enough parallax – *i.e.*, those that are close to the camera and situated perpendicularly to its trajectory – and therefore they need to operate in indoor scenarios with lateral motions. Later, *undelayed landmark initialization* [3] (*i.e.*, mapping the landmarks from their first, partial observation) allowed the inclusion of low parallax landmarks, *i.e.*, those that are remote and/or situated close to the motion axis. This permits operation in outdoor scenes with frontal trajectories. Recently, *Inverse Depth Parameterization* (IDP) [4], [5] has proved to correctly map any landmark ranging from nearby to infinity. Today, we can say that monocular SLAM systems are mature in the sense that they are able to exploit the whole geometrical information coming from the camera.

It is often highlighted by roboticists that the main difference between SFM and monocular SLAM is that the former

Joan Solà is currently with the Artificial Intelligence Center at SRI International, 333 Ravenswood ave, Menlo Park, 94025 CA, USA. Email: sola@ai.sri.com. This work was performed as part of the author's Ph.D. at LAAS-CNRS, France.

is solved offline, via the iterative nonlinear optimization method known as bundle adjustment (BA), while the latter must be causally solved online, thus making use of stochastic estimators or *filters* that naturally provide incremental operation. This is only partially true: the differences between SFM and SLAM are also in the objectives, which means that the different aspects of the problem are given different priorities.

In particular, SFM achieves an exploitation of the whole visual information without the difficulties encountered in monocular SLAM. Let us try to understand this curious fact. SFM considers the structure as a final objective, *i.e.*, as a result of the whole process, and the emphasis is on minimizing the errors in the *measurement space*, therefore making proper use of all the measured information. On the contrary, in SLAM the map must be continuously used to make the operation proceed, and therefore it is given a more central role, with some of the operations (and particularly landmark initialization) being performed in map space, which is the system's *state space*. The higher dimension of this state space (with respect to the measurement space) prevents it from being statically observable, leading to the difficulties mentioned. Informally, we could say that modern undelayed methods for monocular SLAM are almost equivalent to an operation in the measurement space: they initialize the information in the map space partially, *i.e.*, exactly as it comes from the measurement space. A similar point of view for this concept can be found in [6].

B. Stereo-vision SLAM and Visual Odometry (VO)

The ability of a stereo assembly to immediately provide 3D landmark estimates allows us to use them to feed the best available SLAM algorithms. Most such SLAM systems consider the stereo assembly as being a single monolithic sensor, capable of gathering 3D geometrical information [7], [8]. This fact, that seems perfectly reasonable, is the main paradigm questioned in this article because, by considering two linked cameras as a single 3D sensor, the following two drawbacks become very difficult to overcome:

1) Limited 3D estimability range: While cameras are capable of sensing visible objects that are potentially at infinity, a stereo rig provides reasonably good 3D estimates only to a limited range, usually from 3 m to a few tens of meters depending on the baseline. Because (classical, nonmonocular) SLAM algorithms expect full 3D estimates for landmark initialization (*i.e.*, they reason in the map space), only information belonging to this limited region can be used for SLAM.

2) Mechanical fragility: If we want to extend the 3D estimability range we need to a) increase the stereo baseline and b) keep or improve the overall sensor precision. Achieving both needs simultaneously requires delicate, heavy and/or expensive mechanical structures.

One could say that, at least in the methodological aspects, visual odometry (VO) is to stereo SLAM what SFM is to monocular SLAM. Advanced VO solutions achieve very low drift levels by making use of a) advanced real-time image processing [9], b) dense image information [10], or c) bundle adjusting the set of the N oldest frames together with additional fusion with an inertial measurement unit (IMU) [11]. The same considerations made for SFM apply to VO when considering the information it is able to exploit: by using BA methods reasoning in the measurement space, VO is naturally able to exploit all the sensory information.

C. Sensor fusion and Multisensor SLAM

The fact of SLAM being solvable by filtering allows us to envision SLAM systems in the way filters have been used in control theory: as sensor fusion systems. Let us highlight some of the assets of filtering in sensor fusion:

1) Multisensor operation: An unbounded amount of sensors can be operated in a consistent framework.

2) Desynchronized operation: The data rates of all these sensors do not need to be synchronized.

3) Decentralized operation: Advanced filter formulations in information form [12] achieve decentralized operation with delayed or asequent measurements without the need for a central fusion agent.

4) Sensor self-calibration: Unknown sensor parameters can be estimated by the filters provided that they are observable via the overall system's sensor configuration [13].

However, when using several equal sensors for SLAM, we observe a widespread practice of not using fusion: instead, the extrinsic parameters linking the sensors are offline calibrated and the set is then treated as a monolithic supersensor. This is the case of two 180° scanners to simulate a 360° scanner, or the mentioned stereo rig to simulate a 3D sensor. A sensor fusion approach in these cases should naturally provide to the SLAM system all the aforementioned assets.

D. Multicamera SLAM and the aim of this article

The key idea that we want to convey is very simple: by using the central SLAM filter as a fusion engine, we will be able to use any number of cameras with a complete freedom of configuration; by treating them as monocular sensors with modern undelayed initialization methods, we will extract all the available geometrical information provided by the images. The filter will be responsible for, and we will be liberated of, making the 3D properties of the perceived world arise.

Applications may go from the simplest stereo system, through robots with several unequal cameras (a panoramic camera for localization and a perspective camera looking forward for reactive navigation), to multirobot cooperative VSLAM where monocular observations from different robots may be used to determine the 3D locations of very distant landmarks. However, just for the sake of simplicity and without significant loss of generality, this article will treat the case of a robot equipped with a stereo rig. This is the sensor assembly that we consider in what is to follow, and we leave for the reader the exportation of our ideas and techniques to other possible configurations.

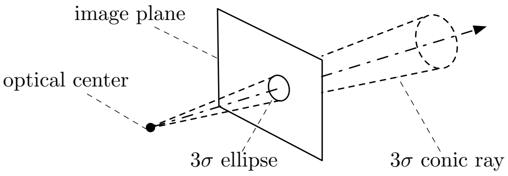

Fig. 1. The conic ray back-projects the elliptic representation of the Gaussian 2D measure. It extends to infinity.

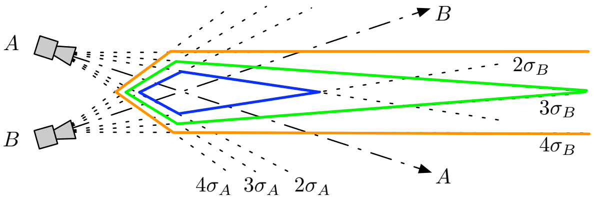

Fig. 2. Different regions of intersection for 4σ (*orange*), 3σ (*green*), and 2σ (*blue*) ray widths when the outer 4σ bounds are parallel. The parallax or angle between ray axes A and B is $\psi = 4\left(\sigma_A + \sigma_B\right)$.

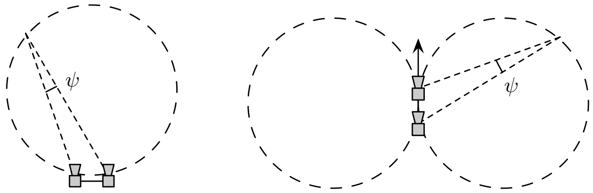

Fig. 3. Simplified depth estimability regions in a stereo rig (*left*) and a camera traveling forward (*right*). The angle ψ assures estimability via different points of view.

III. Monocular SLAM

We present the ideas and material necessary to develop our sensor fusion approach to VSLAM. We introduce the concept of estimability that will help to clarify the key properties of undelayed monocular SLAM, and describe a particular solution to monocular SLAM named FIS-SLAM [3].

A. 3D estimability from vision

When a new feature is detected in the image, the back-projection of its noisy-measured position defines a conic-shaped *pdf* for the landmark position, called *ray*, that extends to infinity (Fig. 1). Let us consider two features extracted and matched in two images, corresponding to the same landmark: their back-projections are the conic rays A and B. Their angular widths can be defined as a multiple of the standard deviations σ_A and σ_B of the angular errors (Fig. 2). We may say that the landmark's depth is fully estimated if the region of intersection of these rays is *a)* closed and *b)* sufficiently small. By considering the case where the two external 4σ bounds of the rays are parallel, we can assure that the 3σ intersection region (98% probability) is closed and that the 2σ one (74%) is small. The depth's sigma-to-mean ratio (a measure of linearization validity in EKF [1], [3]) is better than 0.25. The *parallax* angle ψ between the two rays axes is then $\psi = 4(\sigma_A + \sigma_B) = constant$.

In 2D, we can plot the locus of points at the limit of estimability (Fig. 3), which is circular for constant parallax [14]. The landmark is *fully* estimable inside and *partially* outside. The circle's radius is $R = d/(2\sin\psi)$, directly proportional to the distance d between the two cameras and, for small parallaxes, inversely to the angular uncertainties. In 3D, the full estimability region is obtained by revolution of this circle around the axis joining both cameras, producing something like a torus shape. Beyond the estimability boundaries (remote landmarks in a stereo configuration or frontal landmarks in the monocular case, see Fig. 3) full 3D estimates are not possible. If we want an undelayed exploitation of the full visual information we must initialize these landmarks partially.

In multicamera systems, by combining both monocular and stereo capabilities we get an instant estimability of close frontal objects while still utilizing the information of remote objects: the first beneficiary is the robot localization as we will dispose of long-term absolute angular references.

B. Undelayed monocular FIS-SLAM

The core algorithm we use is Federated Information Sharing SLAM (FIS-SLAM), a bearings-only EKF-based SLAM algorithm we presented in [3], which is briefly described as follows. FIS-SLAM focuses on an undelayed way to initialize landmarks, which at the first observation are only partially observed, and their estimates are therefore conic *pdfs* or rays. To achieve this, each ray is first truncated at minimum and maximum considered depths, and then approximated by a geometric series of Gaussians, that we name *ray members*. The geometric series provides good scaling to big distances with a reduced number of members. All ray members are initialized in an EKF-SLAM map as if they were different landmarks, and are assigned a uniform initial probability or *weight*. As new observations are incorporated, these weights are updated with the current likelihoods. The members are progressively deleted as their weights drop below a certain threshold. The landmark is considered fully 3D estimated when only one member is left. Before this happens, the so-mapped members allow correction steps to be performed on the SLAM map by means of a special EKF-based update scheme, Federated Information Sharing (FIS). Inspired by the Principle of Measurement Reproduction [15], FIS federatively (*i.e.*, nonhomogeneously) shares the information provided by the observation among all members and then applies an EKF update on each of them. Thanks to this possibility of updating the map on reobservation of partially mapped landmarks, low parallax landmarks can contribute to SLAM, and remote landmarks may be used as long-term absolute angular references.

IV. Bi-camera SLAM

We describe now the implementation of all the exposed concepts for the case of an extrinsically decalibrated stereo rig. Details for landmark initialization, map updates, and extrinsic self-calibration are given.

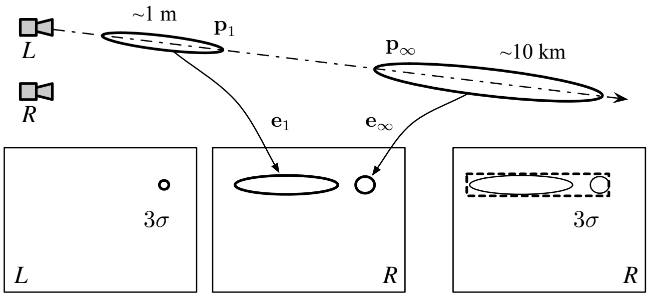

Fig. 4. The 2-member ray in BiCam initialization. The left and right 3σ projections, and the search region in the right image.

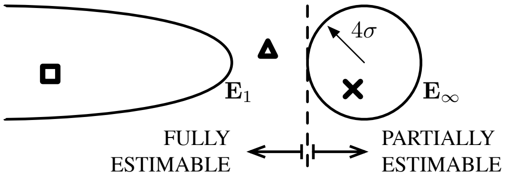

Fig. 5. Deciding on 3D estimability. A 4σ criterion is *a priori* reasoned in the 2D image plane. The measure marked 'cross' corresponds to a landmark outside the stereo estimability region. Landmarks measured 'square' and 'triangle' are inside.

A. Landmark initializations

As a general idea, one should simply initialize landmarks following monocular techniques from the first camera and observe them from the second camera: the fusion filter would determine their 3D positions with more or less accuracy depending on their degree of estimability.

Thanks to the fusion approach, initialization from one camera and the first observation from the second camera do not even need to be synchronized in the same frame instant. However, the particularities of FIS-SLAM suggest proceeding carefully. In FIS-SLAM, mapping depth ranges from very close to very far is costly because the number of ray members is logarithmic with the maximum-to-minimum depths ratio. It is clear that, for nearby landmarks for instance, initializing the whole ray and then deleting all but the right members is not so clever. By understanding the angular properties that generated the estimability regions (Section III-A), we can *a priori* evaluate, from both images, whether or not each landmark is fully 3D-estimable. Landmarks will be initialized in a different manner depending on the result of this evaluation.

Consider two cameras in typical stereo configuration and no distortion. Assume that a new feature $\mathbf{b}_L \sim \{\mathbf{y}_L; \mathbf{R}\}$ is detected in the left image. Its back projection function is $\mathbf{g}(\mathcal{C}_L, s, \mathbf{b}_L)$, with $\mathcal{C}_L$ the left camera pose, s the landmark's depth, and $\mathbf{b}_L$ the bearing, measured at $\mathbf{y}_L$. The feature defines a conic ray for its associated landmark's *pdf*. We define two members $\mathbf{p}_1$ and $\mathbf{p}_\infty$ in this ray (Fig. 4 *top*), one at the minimum and one at the maximum (virtually at infinity) considered depths s_1 and s_∞. Their means and covariances are

$$\bar{\mathbf{p}}_i = \mathbf{g}(\mathcal{C}_L, s_i, \mathbf{y}_L) \tag{1}$$

$$\mathbf{P}_i = \mathbf{G}_{\mathbf{b},i}\,\mathbf{R}\,\mathbf{G}_{\mathbf{b},i}^\top + \mathbf{G}_{s,i}\,\sigma_i^2\,\mathbf{G}_{s,i} \tag{2}$$

with $i \in \{1, \infty\}$, $\sigma_i = \alpha s_i$, $\alpha \leq 0.3$ the shape factor of the geometric ray members, and $\mathbf{G}_\mathbf{b}$ and $\mathbf{G}_s$ the appropriate Jacobian matrices. We project this ray onto the right image: the nearby member becomes an elongated ellipse; the remote member, that projects exactly at the vanishing point of the ray, is a rounded, smaller ellipse (Fig. 4 *bottom*). The axis joining both ellipses centers is the epipolar line, and in the case we are considering it is a horizontal line.

Let $\mathbf{b}_R = \mathbf{h}(\mathcal{C}_R, \mathbf{p})$ be the observation function of the landmark from the right camera. Let $\mathbf{R}$ be the covariance matrix of the right-hand camera measurement noise, and $\mathbf{P}_{\mathcal{C}_R}$ that of this camera's pose uncertainty. The projected $n\sigma$ ellipses correspond to the expectations $\mathbf{e}_i \sim \mathcal{N}\{\bar{\mathbf{e}}_i; \mathbf{E}_i\}$ defined by

$$\bar{\mathbf{e}}_i = \mathbf{h}(\bar{\mathcal{C}}_R, \bar{\mathbf{p}}_i) \tag{3}$$

$$\mathbf{E}_i = \mathbf{H}_{\mathbf{p},i}\,\mathbf{P}_i\,\mathbf{H}_{\mathbf{p},i}^\top + \mathbf{H}_{\mathcal{C},i}\,\mathbf{P}_{\mathcal{C}_R}\,\mathbf{H}_{\mathcal{C},i}^\top + \mathbf{R} \tag{4}$$

with $\mathbf{H}_{\mathbf{p},i}$ and $\mathbf{H}_{\mathcal{C},i}$ the appropriate Jacobians of $\mathbf{h}$.

Following the active feature search approach [16], the region including both 3σ ellipses is scanned for a feature match. The found pixel $\mathbf{y}_R$ is sent to the following 4σ estimability evaluation test (Fig. 5), equivalent to that in Section III-A: *"The measured landmark is considered fully 3D estimable if and only if the measured feature falls strictly at the left-hand side of the* $\mathbf{e}_\infty$ *ellipse's leftmost* 4σ *border."*

If we write the measured pixel $\mathbf{y}_R$ and the remote expectation $\mathbf{e}_\infty \sim \mathcal{N}\{\bar{\mathbf{e}}_\infty; \mathbf{E}_\infty\}$ as

$$\mathbf{y}_R = \begin{bmatrix} y_u \\ y_v \end{bmatrix}, \quad \bar{\mathbf{e}}_\infty = \begin{bmatrix} \bar{e}_{\infty,u} \\ \bar{e}_{\infty,v} \end{bmatrix}, \quad \mathbf{E}_\infty = \begin{bmatrix} \sigma_{\infty,u}^2 & \sigma_{\infty,uv}^2 \\ \sigma_{\infty,uv}^2 & \sigma_{\infty,v}^2 \end{bmatrix},$$

where $(\cdot)_u$ denotes horizontal coordinates, then this criterion resumes simply to

$$y_u < (\bar{e}_{\infty,u} - 4\,\sigma_{\infty,u}) \iff \text{FULLY ESTIMABLE.} \tag{5}$$

The landmark is initialized either as a single Gaussian or as a ray as follows:

1) Fully estimable: We compute the landmark's depth by triangulation, and initialize a 'ray' of one single member at this depth using one of the views. Notice that this depth calculation is used only to provide a good linearization point, not a full 3D estimate. We immediately update it with the second view to refine its position.

2) Partially estimable: A ray is initialized with its closest member already outside the estimability region. The region limit in the ray direction is determined by triangulation with a virtual measurement at the critical point $\mathbf{y}_R^* = [y_u^*, y_v^*]^\top$ with $y_u^* = \bar{e}_{\infty,u} - 4\sigma_{\infty,u}$ and y_v^* chosen so that $\mathbf{y}_R^*$ lies on the epipolar line, that is

$$\mathbf{y}_R^* = \bar{\mathbf{e}}_\infty - \frac{4\sigma_{\infty,u}}{\bar{e}_{1,u} - \bar{e}_{\infty,u}} \cdot (\bar{\mathbf{e}}_{1,u} - \bar{\mathbf{e}}_{\infty,u}). \tag{6}$$

B. Map updates

Thanks to the monocular vision formulation, updates can be performed at any monocular observation of landmarks. This includes any nearby or remote landmark that is visible from both cameras or from only one camera. We use the information-gain driven *active feature search* [16] to select the set of the most valuable landmarks to measure. We measure around 20 features per frame, independently of their character (single Gaussian or ray) and of the camera from which they are observed.

C. Stereo rig extrinsic self-calibration

Stereo rigs are mechanically delicate, especially for large baselines. We believe that stereo assemblies are practical only if they are relatively small ($10 - 20$ cm) or if their main extrinsic parameters are continuously self-calibrated. Outdoor operation may impose this second case, making self-calibration an interesting capability. Notice that the extrinsic parameters are observable: by performing monocular SLAM twice, we are able to obtain the localizations of both cameras with respect to the world. If we associate perceptions on both sides to a unique map, the cameras get localized in the same reference frame, providing the means to self-calibrate the relative transformation between them.

Not all six extrinsic parameters (three for translation, three for orientation) need to be calibrated. In fact, the notion of *self-calibration* inherently requires the system to possess its own gauge. In our case, the metric dimensions or *scale factor* of the whole world-robot system can be obtained only from either the stereo rig baseline, which is one of the extrinsic parameters (and notice that then it is absurd to self-calibrate the gauge!), or from the odometry sensors, which often are much less accurate than any rude measurement we could make of this baseline. Moreover, as cameras are actually angular sensors, vision measurements are much more sensitive to the camera orientations than to any translation parameter. This means that vision measurements will contain little information about these translation parameters. In consequence, self-calibration should concern only orientation, and more precisely, the orientation of the right camera with respect to the left camera. The error of the overall scale factor will be the same as the relative error when measuring the rig's baseline.

We have used a very simple self-calibration solution that has given promising results: we just add three angles (or any other orientation representation we are familiar with – we actually use quaternions) to the EKF-SLAM state vector and let EKF make the rest. The time-evolution function of the extrinsic parameters $\mathbf{q}_R$ is simply $\mathbf{q}_R^+ = \mathbf{q}_R + \gamma$, where γ is a white, Gaussian, low-energy process noise that accounts for eventual decalibrations.

Although it works fairly well, this solution lacks some robustness and is included here as an illustration of the BiCam capability of working with online extrinsic calibration. This fact – this lack of robustness – is observed and analyzed during experiments and further discussed in Section VI.

V. Experiments

The 'whiteboard' indoor experiment is set up as follows. A robot with a stereo head looking forward is run for some 15 m in a straight line inside the robotics lab at the Laboratory for the Analysis and Architecture of Systems (LAAS) (Fig. 6). More than 500 image pairs are taken at approximatively 5 Hz frequency. The robot approaches the objects to be mapped, a situation that is common in mobile robotics but that presents observability difficulties for monocular SLAM because of the singular trajectory. The stereo rig consists of two intrinsically calibrated cameras arranged as indicated in Table I. The left camera is taken as reference, thus deterministically specified, and the orientation of the right camera is initialized with an uncertainty of $1°$ standard deviation. A simple 2D odometry model is used for motion predictions: nominal motion is performed in the robot's local horizontal plane but motion uncertainty is added to all dimensions (three translations and three rotations). This experiment shows the metric accuracy of the resulting map, the self-calibration procedure, and the conditional initialization mechanism.

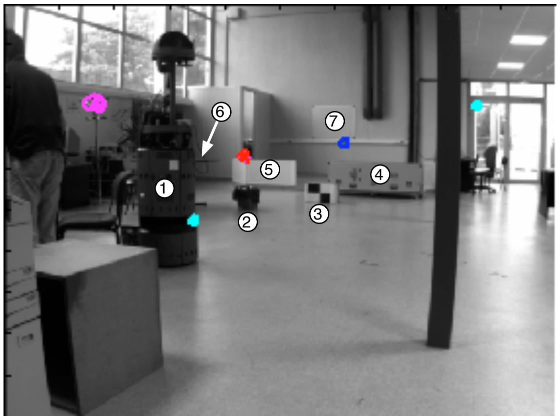

Fig. 6. The LAAS robotics lab. The robot will approach the scene in a straightforward trajectory. We notice in the scene the presence of a robot ①, a bin ②, a box ③, a trunk ④, a fence ⑤, a table ⑥ (hidden by the robot in this image), and the whiteboard ⑦ at the end wall.

TABLE I

Stereo rig parameters in the 'whiteboard' experiment

Scope	Parameters = Values
Dimensions	Base line = 330 mm
Orientation	{pan, tilt} = $\{0°, -5°\}$
Orientation - Euler	$\{\phi, \theta, \psi\}$ = $\{0°, 5°, 0°\}$
Cameras	{resolution,FOV} = $\{512 \times 384 \text{ pix}, 55°\}$
Right camera uncertainties	$\{\sigma_\phi, \sigma_\theta, \sigma_\psi\}$ = $\{1°, 1°, 1°\}$

A. Metric accuracy

We show in Fig. 7 the top-view map of the LAAS robotics lab generated during this experiment. To contrast it against reality, two tests are performed: planarity and metric scale (Fig. 8). 1) The four corners of the whiteboard are taken together with nine other points at the end wall to test coplanarity: the 13 mapped points are found to be coplanar within 4.9 cm of standard deviation error. 2) The lengths of the real and mapped segments marked in red in Fig. 8 are summarized in Table II. The whiteboard has a physical size of 120×90 cm but we take real measures from the

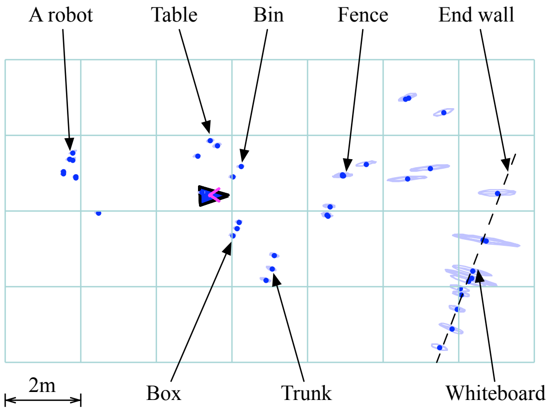

Fig. 7. Map produced by the robot using BiCamSLAM.

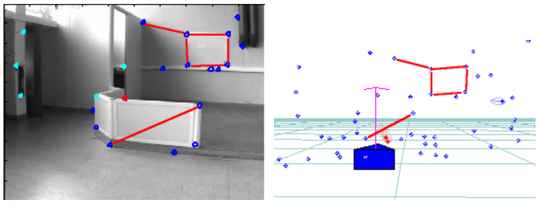

Fig. 8. Metric mapping. The magnitudes of some segments in the real lab are compared to those in the map (*red lines*).

approximated corners where the features are detected as shown in Fig. 8 bottom left and right. We observe errors in the order of 1 cm for landmarks that are still about 4 m away from the robot.

B. Self-calibration

A typical evolution of the three Euler angles representation of the self-calibrated quaternion is illustrated in Fig. 9. We observe the following behavior:

1) Pitch - θ: Pitch angle (cameras tilt $5°$ nominal value) is observable from the first matched landmark. It rapidly converges to an angle of $4.87°$ and remains very stable during the whole experiment.

2) Roll - ϕ: Roll angle is observable after at least two landmarks are observed. It may take some frames for this condition to arrive (here we purposely limited the number of initializations in order to exaggerate this effect) but then it also converges relatively fast and quite stably.

3) Yaw - ψ: Yaw angle is very weakly observable because it is coupled with the landmarks' depths: both yaw angle and landmark depth variations produce a similar effect in the right image, *i.e.*, the feature moves following the landmark's epipolar line. For this reason, yaw starts converging from the first initial uncertainty, but after some frames it does it insecurely and slowly: see how from frame 15 onward yaw uncertainty is already bigger than roll uncertainty, which started converging later. As can be appreciated in Fig. 9 (*right*) the value of the estimated yaw angle shows reasonable convergence only after 150 frames. This is the time monocular SLAM takes to obtain reasonably good 3D estimates of the first perceived landmarks, and therefore the time at which good extrinsic observability is achieved (Section IV-C).

TABLE II
MAP TO GROUND TRUTH COMPARISON.

segment	location	real (cm)	mapped	std. dev.
A	board	119	119.6	0.81
B	board	86	84.3	0.87
C	board	115	114.8	1.11
D	board	88	89.0	0.72
E	wall	134	132.5	0.91
F	fence	125	124.5	1.21

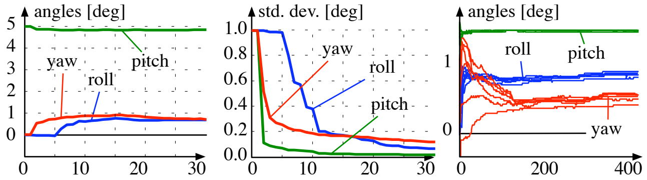

Fig. 9. Extrinsic self-calibration. *Left:* The three Euler angles during the first 30 frames. *Center:* Standard deviations. *Right:* Repeatability of the calibrated angles for 400 frames (pitch not to scale).

TABLE III
SELF-CALIBRATION ERRORS WITH RESPECT TO OFF-LINE CALIBRATION.

Euler angle	calibration		error	std. dev. σ	
	off-line	self-		statistical	estimated
roll ϕ	0.61°	0.60°	-0.01°	0.038°	0.021°
pitch θ	4.74°	4.87°	0.13°	0.006°	0.006°
yaw ψ	0.51°	0.33°	-0.18°	0.108°	$\boxed{0.018°}$

To further investigate yaw stability and consistency, we made 10 runs of 200 frames each and collected the estimated calibration angles and standard deviations at the end of each sequence. We computed the statistical standard deviations (with respect to the 10 runs) of these estimated angles. We compared these values against the angles provided by the Matlab calibration toolbox. Apart from the mentioned initial stability issues, the results in Table III show good calibration, with similar statistical and estimated standard deviations, except for yaw, which shows a clear inconsistency, *i.e.*, an overestimate of its standard deviation (boxed value in the table).

C. Initialization mechanism

The dynamic observability decision criterion with extrinsic self-calibration is illustrated in the sequence of Fig. 10. On the left (right camera images) we see the 3σ search ellipses of expectations $\mathbf{e}_1$ and $\mathbf{e}_\infty$ and on the right a top view of the map. Observe how, in the first frame, extrinsic self-calibration is poor and results in big decision ellipses (decision ellipses are 4σ, thus 33% bigger than the plotted search ellipses), giving initializations of nearby landmarks in the form of rays. Observations from the right camera refine the extrinsic precision, and subsequent decision ellipses become smaller. In the third frame, the stereo rig is already

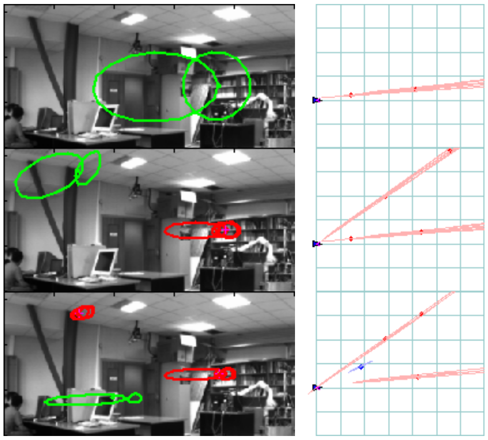

Fig. 10. Landmark initialization in the presence of extrinsic self-calibration. *Left:* The two 3σ ellipses in the right image (*green*). *Right:* The map with rays (partial landmarks) in red and points (full landmarks) in blue. The grid is at 2 m spacing. Observations improve calibration and the ellipses shrink, increasing the boundaries of the 3D estimable region.

quite accurate and is able to fully observe the 3D position of new landmarks (blue landmark in the map).

VI. Conclusions

We have shown that using a sensor fusion approach with monocular SLAM techniques in stereovision or multicamera-equipped robots provides several advantages. These advantages have been highlighted and explored with a particular bearings-only SLAM algorithm applied to an extrinsically decalibrated stereo rig, although they should come up naturally in any other implementation.

The self-calibration solution proposed here suffers from poor observability and inconsistency problems. Theoretically speaking, lack of observability should not be a problem as an image pair of five 3D points in a general configuration renders the whole system observable, but things are in practice much more delicate. Regarding inconsistency, the fact of the different ray members being projected from one camera to the other camera seems to be the responsible of the observed fall in the predicted uncertainty of the yaw angle. This occurs because upon observation of a multihypothesized ray from the right camera, the FIS update [3] may produce overestimated values in the direction where the ray's members' expectations are more dispersed, which is precisely the direction that couples the cameras' convergence angle (the yaw angle) with the distance to the landmarks. To provide a consistent, real-time, continuous calibration operation, we believe the Inverse Depth Parametrization (IDP) in [4] could give more satisfying results. The IDP is additionally able to encode depths up to infinity, therefore providing the conditions to achieve all the assets of the sensor fusion approach that we have just defended. Nevertheless, our procedure helped to prove with real experiments that, given a dynamic extrinsic calibration with its time-varying uncertainty, the 3D observability can be easily determined at every moment from very simple reasoning on the image plane. Of course one can use the whole BiCam proposal with an offline-calibrated stereo rig.

References

[1] A. Davison, "Real-time simultaneous localisation and mapping with a single camera," in *International Conference on Computer Vision*, Nice, October 2003.

[2] T. Bailey, "Constrained initialisation for bearing-only SLAM," *IEEE International Conference on Robotics and Automation*, vol. 2, pp. 1966–1971, 2003.

[3] J. Solà, A. Monin, M. Devy, and T. Lemaire, "Undelayed initialization in bearing only SLAM," in *IEEE International Conference on Intelligent Robots and Systems*, Edmonton, Canada, 2005. [Online]. Available: http://www.laas.fr/~jsola/publications/UndelayedBOSLAM.pdf

[4] J. Montiel, J. Civera, and A. Davison, "Unified inverse depth parametrization for monocular SLAM," in *Robotics: Science and Systems*, Philadelphia, USA, August 2006.

[5] E. Eade and T. Drummond, "Scalable monocular SLAM," *IEEE Int. Conf. on Computer Vision and Pattern Recognition*, vol. 1, pp. 469–476, 2006.

[6] J. Folkesson, P. Jensfelt, and H. I. Christensen, "Vision SLAM in the measurement subspace," in *IEEE Int. Conf. on Robotics and Automation*, Barcelona, 2005.

[7] J. Diebel, K. Reuterswärd, S. Thrun, and R. G. J. Davis, "Simultaneous localization and mapping with active stereo vision," in *IEEE/RSJ International Conference on Intelligent Robots and Systems*, IEEE, Ed., Sendai, Japan, October 2004.

[8] P. Elinas, R. Sim, and J. J. Little, "s-SLAM: Stereo vision SLAM using the Rao-Blackwellised particle filter and a novel mixture proposal distribution," in *IEEE Int. Conf. on Robotics and Automation*, IEEE, Ed., 2006.

[9] T. D. Barfoot, "Online visual motion estimation using FastSLAM with SIFT features," in *IEEE International Conference on Intelligent Robots and Systems*, August 2005.

[10] A. I. Comport, E. Malis, and P. Rives, "Accurate quadrifocal tracking for robust 3D visual odometry," *IEEE International Conference on Robotics and Automation*, pp. 40–45, 2007.

[11] K. Konolige, M. Agrawal, and J. Solà, "Large-scale visual odometry for rough terrain," October 2007, accepted for publication at ISRR07.

[12] E. Nettleton, H. Durrant-Whyte, and S. Sukkarieh, "A robust architecture for decentralised data fusion," in *IEEE International Conference on Advanced Robotics*, 2003.

[13] E. M. Foxlin, "Generalized architecture for simultaneous localization, auto-calibration, and map-building," in *IEEE/RSJ Conf. on Intelligent Robots and Systems*, 2002.

[14] J. Solà, "Towards visual localization, mapping and moving objects tracking by a mobile robot: a geometric and probabilistic approach." Ph.D. dissertation, Institut National Polytechnique de Toulouse, February 2007. [Online]. Available: http://www.laas.fr/~jsola/publications/Thesis.pdf

[15] V. A. Tupysev, "A generalized approach to the problem of distributed Kalman filtering," in *AIAA Guidance, Navigation and Control Conference*, Boston, 1998.

[16] A. J. Davison, "Active search for real-time vision," *International Conference on Computer Vision*, vol. 1, pp. 66–73, 2005.