Joint self-calibration between two separate heterogeneous sensor networks providing range and video measurements

Beatriz Quintino Ferreira, João Pedro Gomes, João Paulo Costeira

Institute for Systems and Robotics, Instituto Superior Técnico, Universidade de Lisboa

beatrizquintino@isr.ist.utl.pt, http://users.isr.ist.utl.pt/~beatrizquintino/index.html

Abstract

We present a novel joint self-calibration method for heterogeneous sensor networks, namely for not necessarily collocated sensors that provide range and video information. To be used to as input for hybrid algorithms that solve the localization problem, the spatial information gained from these measurements has to be expressed in a common reference frame, which we choose as any convenient one for describing the coordinates of anchors in a particular setup. Angular information can be provided by a number of technologies, but the emphasis is on low-cost solutions based on video and fiducial markers such as ARUCO [8]. It is then necessary to translate this to the global reference frame, which would be straightforward if the position and pose of the marker, as well as the relative positions of the range sensor and the camera, were exactly known. Manually determining these parameters is an option, but the process is cumbersome and error prone; e.g., range sensors are not point-like, and distances to the camera should be measured with respect to its focal point, which is usually not accessible as it is located inside the enclosure. We propose a more convenient alternative: an automatic calibration process through a "system identification" procedure. In the proposed procedure the transformation between the reference frames of the separate sensor networks can be obtained using collections of target positions, estimated via a range-only algorithm, and rotations and translations between the camera and the visual markers. We evaluate the proposed method through simulation and in an experimental testbed. Specifically, we show that the true solution was consistently obtained with this algorithm. With real data, the solution was consistent across trials and compatible with the characteristics of our testbed.

Keywords. Calibration, heterogeneous sensor networks, orientation, range.

1 Introduction

There are numerous scenarios in which a single type of sensor cannot frequently guarantee a successful localization estimate. Thus, combining range and vision-based methods to create a hybrid localization scheme can compensate for the limitations and shortcomings of the methods alone. This technical report was developed in the context of hybrid localization schemes which, in contrast to conventional methods that rely on a single sensed variable, fuse noisy range measurements with incident angular information extracted from video, between pairs of nodes, and for which the calibration between the sensor networks, i.e. the assimilation of range and video data, is a pivotal precondition. In networks with more than one type of sensor an accurate separate calibration for each type is normally required. Even with different set-ups, such calibrations result in the same rather complex and challenging problem, whose goal is to localize all the sensors in a global frame. In this vein, a joint calibration procedure allows to translate measurements (defined on their own separate coordinate systems) between the two sensor networks, expressing both classes in a global frame. We introduce a closed form calibration procedure between range and video sensors. More specifically, the wireless sensor network providing range measurements may use sensed variables

such as time of flight (TOF) from an acoustic or an electromagnetic device, or received signal strength indication (RSSI), among others. Moreover, our approach assumes the existence of visual features that can be detected and recognized by a network of cameras.

We succeed at deriving and testing a calibration procedure based solely on pairs of range and orientation measurements as input, whose solution enables accurate position estimation for the localization problem in hybrid frameworks (in both single-source and cooperative paradigms).

1.1 Outline

The remainder of this paper is organized as follows: in section 2 we address and discuss relevant state-of-art works on this topic; section 3 introduces the derivation and formalization of the proposed self-calibration method between range and visual sensor networks; in section 4 simulation and experimental results are presented and, finally, main conclusions are drawn in section 5.

Throughout this document, both scalars and individual position vectors will be represented by lower-case letters. Vectors of concatenated coordinates and matrices will be denoted by boldface lower-case and upper-case, respectively. The superscript $*$ stands for the conjugate transpose and T for the transpose of the given real vector or matrix. $\mathbf{I}_m$ is the identity matrix with dimension $m \times m$ and $\mathbf{1}_m$ the vector of m ones. $\otimes$ represents the Kronecker product, and $\|\mathbf{A}\|_F$ the Frobenius norm of matrix $\mathbf{A}$.

2 Related work

In the following, we present what we think are the most relevant works found in the state-of-the-art directly addressing the self-calibration problem with heterogeneous sensor fusion.

There is a variety of algorithms designed for diverse localization approaches that combine information from different sensors. Focusing on localization methods relying on range and visual sensor fusion, common solutions range from probabilistic methods such as [1] to a recently proposed object oriented visual SLAM (Simultaneous Localization And Mapping) dubbed SLAM++ [2], which uses range measurements from a Kinect depth sensor to perform a dense surface acquisition technique. The latter method comprehends recognition of the scene objects in order to register their pose and the camera-object transformations into a graph, optimized with updates of the observations, used for mapping and estimating the camera and objects pose. The previous method achieves very good results. Notwithstanding, SLAM++ presents some drawbacks since it is very susceptible to local minima, due to the underlying Iterative Closest Point (ICP) registration algorithm, thus requiring a good initialization estimate, and becomes extremely demanding computationally in order to maintain a high performance.

Crocco et al., in [3], propose a closed-form solution to jointly calibrate a network of heterogeneous sensors (video and range sensors). In this fusion approach (that assumes the availability of both range distances to the target and visual information on the target position), a rank constraint for both range and image data is used. Specifically, this rank condition allows to find an initial affine solution via bilinear factorization, as in [4]. The work by Crocco et al. [4] shows that an approximate solution to the network self calibration problem exists by transforming the original nonlinear Least-Squares (LS) cost function minimization in a bilinear matrix formulation. Additionally, the proposed algorithm handles the highly probable case of missing data in the measurements.

In this calibration framework [3], both the positions of the range sensors, cameras and target are unknown, apart from the knowledge of the position of some sensor anchors. Yet, if these anchor positions are not available, this algorithm is still able to provide a solution, an arbitrary 3D rotation and a translation away from the practical solution, using only the metric constraints from the cameras.

The range sensor calibration follows, in short, the procedure in [4], obtaining, in the end, a bilinear formulation. As each sensor is supposed to estimate the distance between itself and the target (for example, using RSSI), a matrix whose entries are the squared of the estimated distances can be built. Identically to what is introduced in [4] regarding the range sensors calibration, eliminating the quadratic terms of the equations, one can reach a set of $(m-1) \times (n-1)$ linear equations, where m and n are the number of range sensors and targets, respectively. Then, a SVD is applied to this bilinear matrix formulation, once again, similarly to what is performed in [4]. The elements of a mixing matrix can be found by solving a LS problem, which gives the correct solution to the range calibration. On the other hand, the video camera calibration consists in computing the parameters of an affine camera and the 3D position of a target, only based on image measurements. The derived form for single target moving observed by a certain number of cameras is a classical SFM problem formulation, known in Computer Vision since it leads to efficient closed-form solutions (for both 3D reconstruction and camera calibration). Such method is based on applying SVD to the aligned to the image centroid measurement matrix, similarly to what is made for the range sensors calibration step. The specificity in this case is the exploration of the constraints imposed by the camera models. The final solution for the image calibration can be found via a LS problem.

Computing the joint closed solution using the range and visual constraints of the heterogeneous sensors is possible because both measurement data share a common subspace, defined by the target position. Therefore, combining the bilinear forms found for both types of measurements in a matrix it is possible to apply a single SVD in order to obtain the first factorization, which can be then upgraded to metric. Crocco et al. note that it might be necessary to perform a normalization of the data to balance the SVD results. The final calibration solution can be found recasting the range and cameras constraints to a LS form so that it is possible to jointly solve both constraints through the pseudo-inverse.

Although the clear relevance of the state-of-the-art found on this line [3,5], the approach proposed in the present work differs from the previous and goes beyond, since it surpasses the limitations present in [3,5]. More specifically, [5] shows a space dimension constraint, as the proposed method can be applied exclusively in 2D, and in [3] there is a limitation of the observations, since observations of the same target points must be performed with the different sensors (range and video). In the work now introduced, however, target nodes roaming in the 3D space will be separately sensed by heterogeneous sensors from two different and separated/uncoupled networks: a wireless sensor network and camera network.

3 Joint calibration

3.1 Problem Formulation

Let us consider two separate heterogeneous sensor networks providing a set of m known reference points (also named anchors) with positions $a_i \in \mathbb{R}^n, i = 1, ..., m$. Of these *a priori* known positions,

the ones whose indices belonging to set $\mathcal{R}$ are ranging sensors, thus providing noisy range measurements to the acoustic/electromagnetic source d_i, whereas those with indices in $\mathcal{T}$ measure bearings u_i to the camera. The elements whose positions are unknown and for which an estimate is sought are called targets. In the scenario assumed in this work each target is composed by/comprises a range sensor, measuring the distances between each target and the available acoustic/electromagnetic anchors (implementation dependent), and a camera which obtains orientations between the target and the available visual anchors. As the targets comprise two different modules it is not physically feasible that they completely overlap, therefore an additional variable, the transformation between the center of the acoustic sensor and the center of the camera, must be taken into account.

We propose estimating the three variables that enable expressing all anchor positions a_i in the same global coordinate system in which the target position is determined. The three unknown variables are the rotation matrix and the translation vector between the coordinate systems of the two sensor networks and, due to the use of two different sensors in the target, the translation vector between the camera focal point and the center of the range sensor.

3.2 Joint calibration procedure derivation

From the existence of the two detached sensor networks, which contrasts with the related work found (more specifically in [3, 5]), emerged the need to calibrate both networks, as a precondition to perform localization. In fact, in [5] it is assumed that the obtained measurements are the absolute angles, since all the anchor nodes are aware of their orientation relative to the global coordinate system. This differs from our scenario, in which the network of sensors measuring angles is detached from the network of ranging sensors. Hence, such assumption cannot be made and a method to relate and determine the rigid transformation between the two is required. The present Subsection details the process developed to achieve such calibration.

To fuse the information conveyed by both sensors, which belong to two different sensor networks, both measurements have to be referenced in a common coordinate system. However, as shown in Figure 3.1, the acoustic anchors form a network detached from the network of the visual features. As each network has its own separate coordinate system, it becomes necessary to define a global referential. Furthermore, Figure 3.1 depicts the target (a device comprising a video camera that can measure transformations to the visual markers attached to an acoustic sensor that can gather range measurements to the wireless sensor nodes) performing a trajectory in the localization scenario which includes both types of anchor nodes and all the translations and rotations that can be measured or should be determined through the joint calibration process.

In this case, the range measurements obtained acoustically will lead to a target position estimate referenced in the coordinate system defined by the anchors of the sensor network. On the other hand, the camera pose, relative to the identified visual features, is determined in a coordinate system defined by the camera.

The method that is explained hereinafter succeeds in determining the rigid transformation (rotation matrix and translation vector) between the two sensor networks from data alone, i.e., using pairs of estimated acoustic target positions and camera positions. Hence, this process is called self-calibration, as there is no need for prior information. The massive flexibility and convenience provided by such assumptions-free process is definitely a decisive advantage regarding the experimental deployment when compared to previously proposed methods.

Considering a set-up similar to the one presented in Figure 3.1 and a target formed by a range sensor attached to a camera, an algebraic method for the auto-calibration is derived. Such method

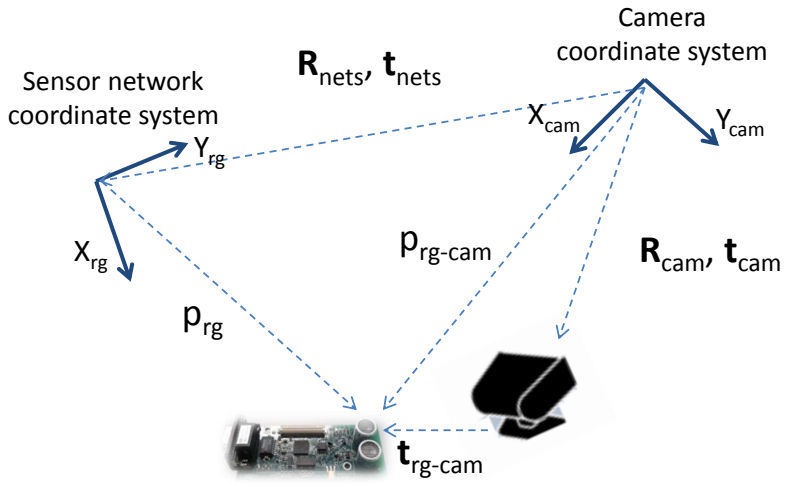

Figure 3.1: Scheme for the self-calibration of the two uncoupled networks (acoustic and visual), and auxiliary to the derivation of expression (3.1)

derivation will use the notation of Figure 3.1 throughout.

The unknowns of this process are $\mathbf{R}_{\text{nets}}$, which denotes the rotation matrix between the camera and the sensor network frames, $\mathbf{t}_{\text{nets}}$, which represents the translation vector between the same two networks, and, since the range and orientation are not measured by the same sensor, it is also necessary to determine $\mathbf{t}_{\text{rg-cam}}$, which is the translation between the focal point of the camera and the center of the range sensor receiver. Additionally, let $\mathbf{R}_{\text{cam}}$ and $\mathbf{t}_{\text{cam}}$ be the rotation and translation, respectively, measured by the camera relative to its own frame and p_{rg} be the range sensor localization computed using the ranges, in the referential defined by the sensor network.

In the envisaged procedure, sets of range measurements, between the anchors and the target, as well as pairs of $\mathbf{R}_{\text{cam}}$ and $\mathbf{t}_{\text{cam}}$, from the camera detecting a visual feature, can be collected. With the first set of data, the position of the range sensor in the sensor network reference system (p_{rg}) can be estimated using an algorithm based on a Least-Squares methodology such as the Squared-Range Least-Squares (SR-LS) [6] or on a Semidefinite Relaxation (SDR) approach, such as the Source Localization in The Complex Plane (SLCP) or the Source Localization with Nuclear Norm (SLNN) for 3D [7]. On the other hand, it is possible to determine the rotation and translation of a camera relative to a visual feature ($\mathbf{R}_{\text{cam}}$ and $\mathbf{t}_{\text{cam}}$), and therefore the camera position in its own coordinate system is straightforwardly computed.

From the assumed localization scenario represented in the scheme of Figure 3.1 follows the relation in equation (3.1), where the notation $p_{\text{rg}}^{(t)}$ is used to represent a time-dependent variable.

$$\begin{cases} p_{\text{rg}}^{(\text{t})} = \mathbf{R}_{\text{nets}} p_{\text{rg-cam}}^{(\text{t})} + \mathbf{t}_{\text{nets}} \\ p_{\text{rg-cam}}^{(\text{t})} = \mathbf{R}_{\text{cam}}^{(\text{t})} \mathbf{t}_{\text{rg-cam}} + \mathbf{t}_{\text{cam}}^{(\text{t})} \end{cases} \tag{3.1}$$

Taking the difference between two generic time instants, t_k and t_0, for the second equation of the system (3.1), one can obtain the expression (3.2):

$$\left[\mathbf{R}_{\text{nets}}(\mathbf{R}_{\text{cam}}^{(\text{t}_\text{k})}\mathbf{t}_{\text{rg-cam}} + \mathbf{t}_{\text{cam}}^{(\text{t}_\text{k})}) + \mathbf{t}_{\text{nets}}\right] - \left[\mathbf{R}_{\text{nets}}(\mathbf{R}_{\text{cam}}^{(\text{t}_0)}\mathbf{t}_{\text{rg-cam}} + \mathbf{t}_{\text{cam}}^{(\text{t}_0)}) + \mathbf{t}_{\text{nets}}\right] = p_{\text{rg}}^{(\text{t}_\text{k})} - p_{\text{rg}}^{(\text{t}_0)}. \tag{3.2}$$

The expression (3.2) can be simplified into (3.3), according to the auxiliary scheme in Figure

3.2, which has the same notation of the algebraic derivation performed below.

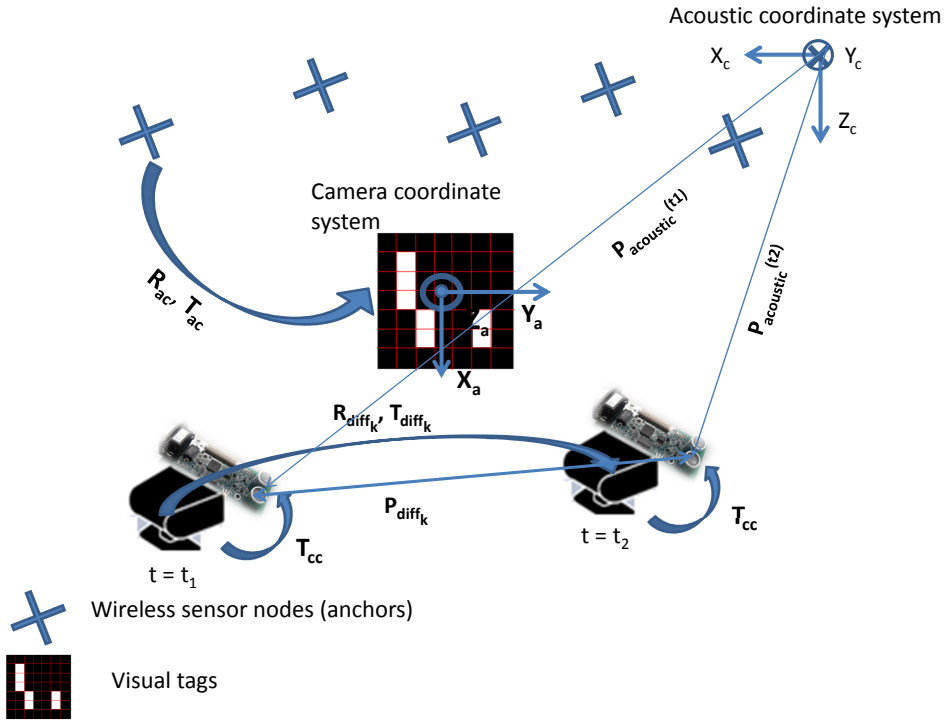

Figure 3.2: Auxiliary scheme for the derivation of equation (3.3)

$$\underbrace{p_{\text{rg}}^{(\text{t}_\text{k})} - p_{\text{rg}}^{(\text{t}_0)}}_{p_{\text{diff}_\text{k}}} = \mathbf{R}_{\text{nets}} \left[\underbrace{(\mathbf{R}_{\text{cam}}^{(\text{t}_\text{k})} - \mathbf{R}_{\text{cam}}^{(\text{t}_0)})}_{\mathbf{R}_{\text{diff}_\text{k}}} \mathbf{t}_{\text{rg-cam}} + \underbrace{(\mathbf{t}_{\text{cam}}^{(\text{t}_\text{k})} - \mathbf{t}_{\text{cam}}^{(\text{t}_0)})}_{\mathbf{t}_{\text{diff}_\text{k}}} \right] \tag{3.3}$$

In (3.3) the translation vector between the sensor network and the camera coordinate systems ($\mathbf{t}_{\text{nets}}$) is canceled, leaving just two unknown variables $\mathbf{R}_{\text{nets}}$ and $\mathbf{t}_{\text{rg-cam}}$.

Horizontally stacking all the N (number of measurements -1) pairwise differences measured, equation (3.3) can be expressed as in (3.4)

$$\underbrace{[p_{\text{diff}_1} \cdots p_{\text{diff}_\text{N}}]}_{\mathbf{p}_{\text{diff}}} = \mathbf{R}_{\text{nets}} \left(\underbrace{[\mathbf{R}_{\text{diff}_1} \mathbf{t}_{\text{rg-cam}} \cdots \mathbf{R}_{\text{diff}_\text{N}} \mathbf{t}_{\text{rg-cam}}]}_{\mathbf{R}_{\text{diff}}(\mathbf{I}_N \otimes \mathbf{t}_{\text{rg-cam}})} + \underbrace{[\mathbf{t}_{\text{diff}_1} \cdots \mathbf{t}_{\text{diff}_\text{N}}]}_{\mathbf{t}_{\text{diff}}} \right), \tag{3.4}$$

where $\mathbf{I}_N$ denotes an identity matrix of $N \times N$, $\mathbf{R}_{\text{diff}}$ is formed by horizontally stacked differences of rotations matrices (which no longer hold the properties of a rotation matrix) and $\mathbf{t}_{\text{diff}}$ includes the pairwise differences of translation vectors.

According to (3.4), the error function can be defined as

$$\mathbf{E} = \left[\mathbf{p}_{\text{diff}} - \mathbf{R}_{\text{nets}} \left(\mathbf{R}_{\text{diff}} (\mathbf{I}_N \otimes \mathbf{t}_{\text{rg-cam}}) + \mathbf{t}_{\text{diff}} \right) \right]. \tag{3.5}$$

At this point we note that solving (3.5) for each of the unknowns, both $\mathbf{R}_{\text{nets}}$ and $\mathbf{t}_{\text{rg-cam}}$ can be independently determined through a closed form solution: either by a Procrustes or a Pseudo-

inverse problem whether $\mathbf{t}_{\text{rg-cam}}$ or $\mathbf{R}_{\text{nets}}$, respectively, is known. Thus minimizing the error in (3.5) it is possible to estimate both variables, as detailed below.

In fact, $\mathbf{R}_{\text{nets}}$ can be estimated by solving an orthogonal Procrustes problem[1], which determines the closest orthogonal rotation matrix that best fits two sets of points. Here, one set of points is given by $\mathbf{p}_{\text{diff}}$ (which are the pairwise differences of the position estimations obtained through the sensor network, alone) and the other by $\mathbf{R}_{\text{diff}}(\mathbf{I}_N \otimes \mathbf{t}_{\text{rg-cam}}) + \mathbf{t}_{\text{diff}}$, hence the solution of the Procrustes problem will give an estimate of $\mathbf{R}_{\text{nets}}$. In order to reduce the complexity of the algebraic calculations in cases such as this, the Frobenius norm is usually applied. Since this norm can be defined as $\|\mathbf{A}\|_F^2 = tr(\mathbf{A}^*\mathbf{A})$, we apply the trace operator to $\mathbf{E}^*\mathbf{E}$, thus $tr\{\mathbf{E}^*\mathbf{E}\}$. Taking the trace truly simplifies the calculations, since the expression is transformed in an inner-product (element-wise computation), while no information is lost, as the eigenvalues hold all the matrix information. This can also be regarded as computing the average of the error covariance matrix. After some algebraic manipulation, a new expression for the error function is obtained in (3.6)[2].

$$\begin{aligned}\mathbf{E}(\mathbf{t}_{\text{rg-cam}}, \mathbf{R}_{\text{nets}}) = tr\{&\mathbf{p}_{\text{diff}}^T\mathbf{p}_{\text{diff}} - 2\mathbf{p}_{\text{diff}}^T\mathbf{R}_{\text{nets}}\left(\mathbf{R}_{\text{diff}}\mathbf{I}_N \otimes \mathbf{t}_{\text{rg-cam}} + \mathbf{t}_{\text{diff}}\right) + \\ &(\mathbf{I}_N \otimes \mathbf{t}_{\text{rg-cam}})^T\mathbf{R}_{\text{diff}}^T\mathbf{R}_{\text{diff}}(\mathbf{I}_N \otimes \mathbf{t}_{\text{rg-cam}}) + 2\mathbf{t}_{\text{diff}}^T\mathbf{R}_{\text{diff}}\mathbf{I}_N \otimes \mathbf{t}_{\text{rg-cam}} + \mathbf{t}_{\text{diff}}^T\mathbf{t}_{\text{diff}}\}\end{aligned} \tag{3.6}$$

To determine the unknown $\mathbf{t}_{\text{rg-cam}}$, we should minimize the error, thereby, using the function $\mathbf{E}(\mathbf{t}_{\text{rg-cam}}, \mathbf{R}_{\text{nets}})$, defined in (3.6) to form the cost function, yields the following optimization problem in (3.7).

$$\begin{aligned}\underset{\mathbf{R}_{\text{nets}}, \mathbf{t}_{\text{rg-cam}}}{\text{minimize}} \quad & tr\{\mathbf{p}_{\text{diff}}^T\mathbf{p}_{\text{diff}} - 2\mathbf{p}_{\text{diff}}^T\mathbf{R}_{\text{nets}}\left(\mathbf{R}_{\text{diff}}\mathbf{I}_N \otimes \mathbf{t}_{\text{rg-cam}} + \mathbf{t}_{\text{diff}}\right) + (\mathbf{I}_N \otimes \mathbf{t}_{\text{rg-cam}})^T\mathbf{R}_{\text{diff}}^T\mathbf{R}_{\text{diff}}(\mathbf{I}_N \otimes \mathbf{t}_{\text{rg-cam}}) \\ & + 2\mathbf{t}_{\text{diff}}^T\mathbf{R}_{\text{diff}}\mathbf{I}_N \otimes \mathbf{t}_{\text{rg-cam}} + \mathbf{t}_{\text{diff}}^T\mathbf{t}_{\text{diff}}\} \\ \text{subject to} \quad & \mathbf{R}_{\text{nets}}^T\mathbf{R}_{\text{nets}} = \mathbf{I}_3\end{aligned} \tag{3.7}$$

The problem in (3.7) is a constrained optimization problem, since the matrix $\mathbf{R}_{\text{nets}}$ has to belong to the manifold of rotation matrices. Consequently, the method to minimize this objective function consists in using the Lagrange multipliers so that this problem is converted into an unconstrained optimization problem (the restriction is incorporated in the same equation as the cost function). The auxiliary function given by the Lagrangian method is shown in (3.8)

$$\mathcal{L}(\mathbf{t}_{\text{rg-cam}}, \mathbf{R}_{\text{nets}}, \boldsymbol{\lambda}) = \mathbf{E}(\mathbf{t}_{\text{rg-cam}}, \mathbf{R}_{\text{nets}}) + tr(\boldsymbol{\lambda}(\mathbf{R}_{\text{nets}}^T\mathbf{R}_{\text{nets}} - \mathbf{I}_3)), \tag{3.8}$$

where $\boldsymbol{\lambda}$ is the matrix with the Lagrange multipliers for all restrictions.

Taking the partial derivative of (3.8) with respect to $\mathbf{t}_{\text{rg-cam}}$ and setting it to zero we can obtain the following expression to determine the unknown $\mathbf{T}_{\text{cc}}$:

$$\mathbf{T}_{\text{rg-cam}} = \left(\sum_i \mathbf{R}_{\text{diff}_\text{i}}^T\mathbf{R}_{\text{diff}_\text{i}}\right)^{-1}\left(\sum_i \mathbf{R}_{\text{diff}_\text{i}}^T\mathbf{R}_{\text{nets}}^T\mathbf{p}_{\text{diff}_\text{i}} - \sum_i \mathbf{R}_{\text{diff}_\text{i}}^T\mathbf{t}_{\text{diff}_\text{i}}\right). \tag{3.9}$$

Additionally, the translation vector between both networks' reference systems $\mathbf{t}_{\text{nets}}$ is computed based on the previously determined unknowns ($\mathbf{R}_{\text{nets}}$ and $\mathbf{t}_{\text{rg-cam}}$) and averaging the time dependent measurements, according to:

$$\mathbf{t}_{\text{nets}} = \overline{\mathbf{p}_{\text{diff}}} - \mathbf{R}_{\text{nets}}\left(\overline{\mathbf{R}_{\text{cam}}}\mathbf{t}_{\text{rg-cam}} + \overline{\mathbf{t}_{\text{cam}}}\right). \tag{3.10}$$

[1] http://en.wikipedia.org/wiki/Orthogonal_Procrustes_problem (accessed in October 2014)

[2] As all the values are real (positions in the 3D space and rotation matrices), the conjugate transpose notation (X^*) is replaced, simply, by the transpose (X^T).

In light of the above algebraic derivation of the formulation to self-calibrate two separate sensor networks, an iterative method to estimate both $\mathbf{R}_{\text{nets}}$ and $\mathbf{t}_{\text{rg-cam}}$ can be performed very efficiently.

Algorithm 1 summarizes the steps of the newly proposed self-calibration procedure.

Algorithm 1 Self-calibration procedure between the acoustic and visual sensor networks

Require: p_{rg}, $\mathbf{R}_{\text{cam}}$, $\mathbf{t}_{\text{cam}}$;
Ensure: $\mathbf{R}_{\text{nets}}$, $\mathbf{t}_{\text{nets}}$, $\mathbf{t}_{\text{rg-cam}}$;
1: $k = 0$
2: Initialize $\mathbf{t}_{\text{rg-cam}} = 0$, $err_{\mathbf{R}_{\text{nets}}} = \text{inf}$;
3: **while** $err_{\mathbf{R}_{\text{nets}}} > thresh$ **do**
4: $k = k + 1$;
5: $\mathbf{R}_{\text{nets}}(k)$ is computed solving the Procrustes problem with p_{diff} and $\mathbf{R}_{\text{diff}}(\mathbf{I}_N \otimes \mathbf{t}_{\text{rg-cam}}) + \mathbf{t}_{\text{diff}}$ as the two sets of points;
6: $\mathbf{t}_{\text{rg-cam}}(k) = \left(\sum_i \mathbf{R}_{\text{diff}_\text{i}}^T \mathbf{R}_{\text{diff}_\text{i}}\right)^{-1} (\sum_i \mathbf{R}_{\text{diff}_\text{i}}^T \mathbf{R}_{\text{nets}}(k)^T p_{\text{diff}_\text{i}} - \sum_i \mathbf{R}_{\text{diff}_\text{i}}^T \mathbf{t}_{\text{cam}_\text{i}})$;
7: $err_{\mathbf{R}_{\text{nets}}} = \|\mathbf{R}_{\text{nets}}(k-1) - \mathbf{R}_{\text{nets}}(k)\|_F$
8: **end while**
9: $\mathbf{t}_{\text{nets}} = \overline{p_{\text{diff}}} - \mathbf{R}_{\text{nets}}(k) \left(\overline{\mathbf{R}_{\text{cam}}} \mathbf{t}_{\text{rg-cam}}(k) + \overline{\mathbf{t}_{\text{cam}}}\right)$;
10: **return** $\mathbf{R}_{\text{nets}} = \mathbf{R}_{\text{nets}}(k)$, $\mathbf{t}_{\text{rg-cam}} = \mathbf{t}_{\text{rg-cam}}(k)$, $\mathbf{t}_{\text{nets}}$;

Since $\mathbf{t}_{\text{rg-cam}}$ is considerably smaller when compared with $\mathbf{t}_{\text{cam}}$, a first estimate for $\mathbf{R}_{\text{nets}}$ ($\mathbf{R}_{\text{nets}}(k = 1)$) can be obtained neglecting $\mathbf{t}_{\text{rg-cam}}$ ($\mathbf{t}_{\text{rg-cam}} = 0$) or, alternatively, with an initialization of the value of $\mathbf{t}_{\text{rg-cam}}$, if available.

4 Results

In this subsection both simulation and real world experiment results validating the proposed self-calibration method are presented.

4.1 Simulation Results

Prior to performing the self-calibration in an experimental set-up, the devised method introduced in Section 3 was validated, both with entirely numerical data and with numerical data for the ranges mixed with real data for the visual information (the notation used below is the same as introduced in 3.2).

For the first case of only simulated data, a fixed translation between the two sensors ($\mathbf{t}_{\text{rg-cam}}$) and translation vectors and rotation matrices between the acoustic and the visual networks ($\mathbf{R}_{\text{nets}}$ and $\mathbf{t}_{\text{nets}}$) were randomly generated in a [0, 10] $\times$ [0, 10] $\times$ [0, 10] cube. For the k-th observation random camera positions ($\mathbf{R}_{\text{cam}_\text{k}}$ and $\mathbf{T}_{\text{cam}_\text{k}}$) were synthesized so that acoustic node positions could be computed, according to the model $p_{\text{rg}_\text{k}} = (\mathbf{R}_{\text{nets}} (\mathbf{R}_{\text{cam}_\text{k}} \mathbf{t}_{\text{rg-cam}} + \mathbf{t}_{\text{cam}_\text{k}}) + \mathbf{t}_{\text{nets}}) (1 + w)$, where $w \sim \mathcal{N}(0, \eta^2)$. Given an initialization for $\mathbf{t}_{\text{rg-cam}}$ (which is admissible as its dimension tends to be an order of magnitude smaller than the other two unknowns), the self-calibration method iterates through estimating $\mathbf{R}_{\text{nets}}$ and $\mathbf{t}_{\text{rg-cam}}$, until the difference between two consecutive estimates of $\mathbf{R}_{\text{nets}}$ is smaller than a predefined threshold. When the latter stopping criterion is verified, $\mathbf{t}_{\text{nets}}$ is computed, using the previously obtained $\mathbf{R}_{\text{nets}}$ and $\mathbf{t}_{\text{rg-cam}}$ estimates. Performing 1000 Monte Carlo

runs of such procedure, making $k = 100$ observations and noise factor 0.01, the errors obtained for the estimated unknowns, computed as $\mathbf{E}_{\mathbf{R}_{\text{nets}}} = \frac{1}{MC}\sum_{i=1}^{MC} \|\mathbf{R}_{\text{nets}_\text{i}} - \mathbf{R}_{\text{netsGT}_\text{i}}\|_F$ for the rotation matrix and $\mathbf{E}_{\mathbf{t}} = \frac{1}{MC}\sum_{i=1}^{MC} \|\mathbf{t}_i - \mathbf{t}_{\text{GT}_\text{i}}\|$ for the translation vectors, were found to be:

$$\mathbf{E}_{\mathbf{R}_{\text{nets}}} = 0.0076 \quad \mathbf{E}_{\mathbf{t}_{\text{nets}}} = 0.0448 \text{ m} \quad \mathbf{E}_{\mathbf{t}_{\text{rg-cam}}} = 0.1810 \text{ m}.$$

The previous results validate the derived self-calibration method, obtaining highly accurate estimates.

Furthermore, to better model this process to the real case, the self-calibration procedure was performed with camera positions ($\mathbf{R}_{\text{cam}_\text{k}}$ and $\mathbf{t}_{\text{cam}_\text{k}}$) acquired in the experimental set-up, using Augmented Reality Tags (ART) from the ARUCO library [8]. Again, 1000 Monte Carlo runs were executed with $\eta = 0.01$, and the results of the simulations with such partial real data were as follows:

$$\mathbf{E}_{\mathbf{R}_{\text{nets}_{\text{half-real}}}} = 0.1027 \quad \mathbf{E}_{\mathbf{t}_{\text{nets}_{\text{half-real}}}} = 0.3361 \text{ m} \quad \mathbf{E}_{\mathbf{t}_{\text{rg-cam}_{\text{half-real}}}} = 0.4359 \text{ m}.$$

The latter results show lower accuracy when compared to the totally simulated case. This may be partially explained by the total number of observations, since simulation with half experimental data comprised 28 pairs of observations and decreasing the number of observations leads to a degeneration of the estimations (the Procrustes problem suffers from ill approximations for small datasets).

4.2 Experimental Results

The experimental set-up devised to deploy the introduced sel-calibration procedure consisted of a wireless network of acoustic sensors (Cricket modules [9]), attached to the ceiling of a room of approximately $50m^3$ (Figure 4.1), and a set of visual features (ART from ARUCO library, in this specific implementation) spread in the same space (Figure 4.2) which can be detected and recognized by a camera. The target, equipped with a Cricket receiver and a camera, can roam in this room, as shown also in Figure 4.2.

Hence, it is possible to determine, for this deployment, the transformation between the two detached networks ($\mathbf{R}_{\text{nets}}$ and $\mathbf{t}_{\text{nets}}$) as well as the translation between the acoustic sensor and the camera of the target ($\mathbf{t}_{\text{rg-cam}}$), using only acquired data of sensor and camera positions during a walk through the set-up.

Both ranges measured between the acoustic anchors and the target sensor and rotations and translations from the visual tags detection (illustrated in Figure 4.3) are collected. Single-source localization methods found in the state-of-the art, such as the SLNN [7] or the SR-LS [6] can be applied to estimate the unknown position of the target acoustic sensor from the range information (in this case we used the SR-LS approach). As for the information relative to the camera, ARUCO library provides the rotation and the translation of the camera relative to the detected markers (respectively $\mathbf{R}_{\text{cam}}$ and $\mathbf{t}_{\text{cam}}$), which are used to determine the camera positions. The acoustic sensor positions estimated in the acoustic sensor network coordinate system (p_{rg}) and the camera positions estimated in the camera coordinate system ($\mathbf{t}_{\text{cam}}$) for a self-calibration test conducted in the experimental set-up are depicted in Figures 4.5 and 4.4. Some target position estimations (namely positions number 9 and 10), seen in Figure 4.5, are clearly outliers, probably due to exceedingly noisy range measurements. Nevertheless, as it will be shown hereinafter, the outcomes of the self-calibration procedure are robust to the presence of such position estimates.

Figure 4.1: Sensor network anchors attached to the ceiling, in the experimental set-up

Figure 4.2: Augmented reality tags spread in the experimental set-up

Figure 4.4 also depicts the map created for the board of ARUCOS on the wall. The origin of the camera referential was defined to be the center of the tag with ID 0. The Cartesian coordinates forming the acoustic anchors constellation were obtained from centering and averaging matrix $\mathbf{D}$ (an Euclidean Distance Matrix (EDM) consisting of all the pairwise Euclidean distances measured between the acoustic sensor nodes), resulting in $\tilde{\mathbf{D}}$ and then applying a single Value Decomposition (SVD) to this new projected and averaged matrix ($\tilde{\mathbf{D}} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^*$). The reconstructed beacon coordinates ($\mathbf{x}$) are computed as $\mathbf{x} = \mathbf{U}\mathbf{\Sigma}^{\frac{1}{2}}$. Later, this constellation will have to be synchronized,

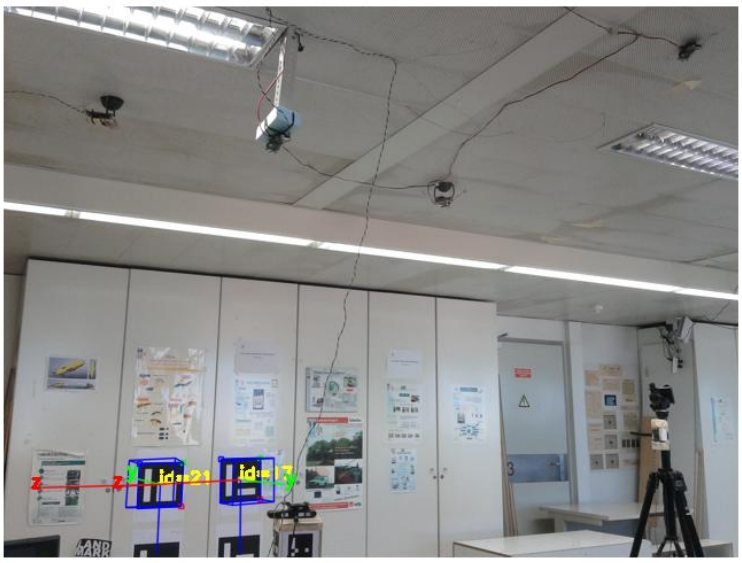

Figure 4.3: Detection of visual features in the heterogeneous sensor networks scenario

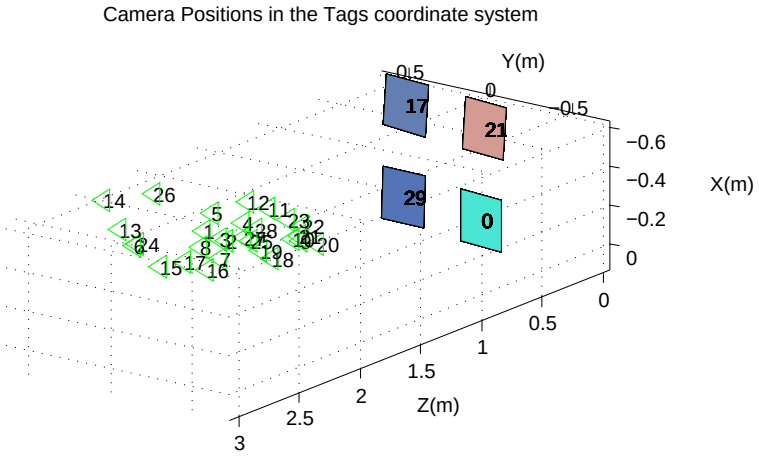

Figure 4.4: Camera positions in the tags coordinate system, collected for the self-calibration input

so that the results can be evaluated.

Taking pairwise differences of both target and camera position estimates results in vectors $\mathbf{p}_{\text{diff}}$ and $\mathbf{T}_{\text{diff}}$, respectively. Additionally, the differences of rotation between the camera and the tags ($\mathbf{R}_{\text{diff}}$) are also determined from the obtained rotations of the ARUCO library. The totality of the inputs of the self-calibration process is, then, computed, and hence the transformation between the two sensor networks ($\mathbf{R}_{\text{nets}}$ and $\mathbf{T}_{\text{nets}}$) and the translation between the target acoustic sensor and camera ($\mathbf{t}_{\text{rg-cam}}$) can be readily estimated for this experimental set-up.

In order to maximize the accuracy of $\mathbf{R}_{\text{nets}}$ and $\mathbf{t}_{\text{nets}}$ estimates in this practical case, and since $\mathbf{t}_{\text{rg-cam}}$ was easily available, the value measured with a tap measure between the acoustic sensor and the focal point of the camera of the target was assigned to $\mathbf{t}_{\text{rg-cam}} = \begin{bmatrix} 0.0800 & -0.0435 & -0.0180 \end{bmatrix}$m.

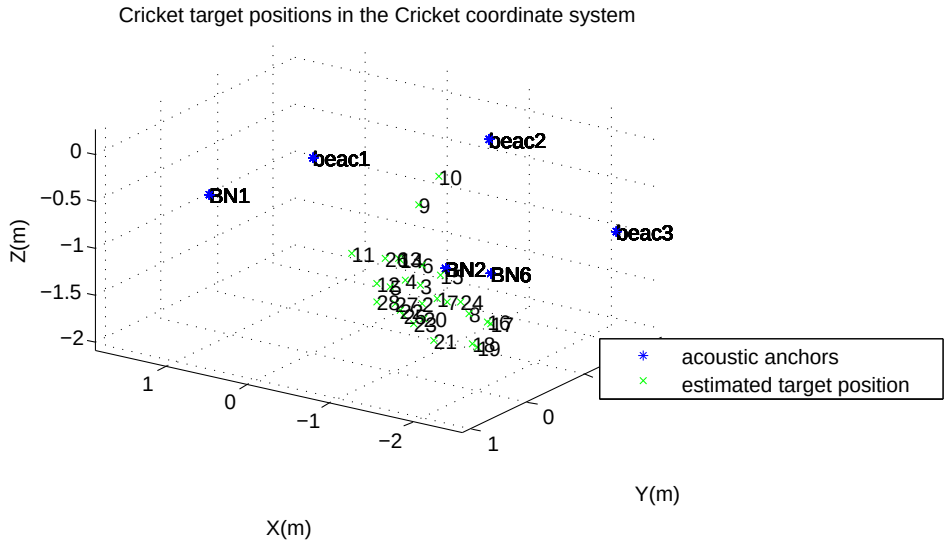

Figure 4.5: Cricket target positions in the Cricket coordinate system, collected for the self-calibration input

The analysis and validation of the achieved results for $\mathbf{R}_{\text{nets}}$ and $\mathbf{t}_{\text{nets}}$ proved to be difficult, as no ground truth or intuitive reference could be defined. Therefore, in order to facilitate the evaluation of the outcomes, the acoustic sensor network referential was aligned with the visual network referential, by performing the synchronization of the aforementioned beacon constellation. For this purpose, one of the acoustic beacons was chosen to be the origin of the coordinate system and the coordinates of other 3 beacons, relative to the new referential (defined in order to be aligned with the origin of the visual tags coordinate system), were measured. Such measured coordinates were taken as reference and used as input, alongside the same 4 sensor positions in the previous non-aligned referential, to solve a Procrustes problem. In this way, rotating all the constellation with the resultant rotation matrix leads to the alignment of the sensor network coordinate system, and with the target sensor positions estimates ($p_{\text{rg-cam}}$) obtained in this new aligned referential, the $\mathbf{R}_{\text{nets}}$ and $\mathbf{t}_{\text{nets}}$ computed can now be compared with a reference and thus evaluated.

For this case, the computed rotation and translation between the sensor network and the visual tags coordinate systems were

$$\mathbf{R}_{\text{nets}} = \begin{bmatrix} -0.0241 & -0.9977 & 0.0640 \\ 0.0941 & -0.0660 & -0.9934 \\ 0.9953 & -0.0179 & 0.0954 \end{bmatrix}, \quad \mathbf{t}_{\text{nets}} = \begin{bmatrix} 0.7495 & 2.0781 & 1.8151 \end{bmatrix} \text{ m}.$$

If the two coordinate systems are considered to be perfectly aligned (which is an assumption that cannot be completely confirmed), $\mathbf{R}_{\text{nets}}$ can be obtained by inspection. Moreover, $\mathbf{t}_{\text{nets}}$ was measured manually with a laser distance meter. These two quantities ground truths were found to be:

$$\mathbf{R}_{\text{netsGT}} = \begin{bmatrix} 0 & -1 & 0 \\ 0 & 0 & -1 \\ 1 & 0 & 0 \end{bmatrix}, \quad \mathbf{t}_{\text{netsGT}} = \begin{bmatrix} 0.82 & 1.97 & 1.84 \end{bmatrix} \text{ m}.$$

This alignment assumption (a physical limitation) is an important issue since a small misalignment

causes changes in the rotation matrix, introducing high sensitivity to the system (as the orientations used for all the position estimates are multiplied by $\mathbf{R}_{\text{nets}}$).

To evaluate the error of the first self-calibration $\mathbf{R}_{\text{nets}}$ estimate, the Frobenius norm of the difference between the experimentally obtained and the ground truth rotation matrices was computed:

$$\mathbf{E}_{\mathbf{R}_{\text{nets}}} = \|\mathbf{R}_{\text{nets}} - \mathbf{R}_{\text{netsGT}}\|_F = 0.1655. \tag{4.1}$$

Moreover, the error of the resultant $\mathbf{t}_{\text{nets}}$ (computed as $\|\mathbf{t}_{\text{nets}} - \mathbf{t}_{\text{netsGT}}\|$) is 0.1314 m.Although such results can be further improved, since they derive from noisy range measurements from the Cricket system, they demonstrate that the derived self-calibration method is correct.

An improvement to this calibration results was accomplished by obtaining the target acoustic sensor position estimates (p_{rg}) from directly measuring the ranges between the target sensor to the anchors with a laser distance meter and not synchronizing the Cricket constellation (as the aligning rotation can introduce additional error, due to possible errors in the reference position measurements). The results obtained with this procedure were

$$\mathbf{R}_{\text{nets}_{\text{laser}}} = \begin{bmatrix} -0.0788 & -0.9955 & 0.0517 \\ 0.0292 & -0.0541 & -0.9981 \\ 0.9965 & -0.0772 & 0.0333 \end{bmatrix}, \quad \mathbf{t}_{\text{nets}_{\text{laser}}} = \begin{bmatrix} 0.7405 & 1.9884 & 1.8310 \end{bmatrix} \text{ m},$$

reflecting an improvement of the estimate of 15% for $\mathbf{R}_{\text{nets}}$, as $\mathbf{E}_{\mathbf{R}_{\text{nets}_{\text{laser}}}}$ is now 0.1406. Likewise, the error norm of the translation vector estimation using the laser is smaller by 37.5%, when compared with the first estimate, as the error norm obtained for this case is 0.0821 m.

This procedure was performed with the intention of not jeopardizing the ensuing hybrid localization performance due to a less accurate calibration. It is emphasised that this does not, in any way, undermine the calibration process as previously proposed, as a self-calibration can always be performed using the Cricket system; however, it is natural that the achieved accuracy is somewhat lower.

5 Conclusions

References

[1] B.-D. Yum, Y.-J. Lee, J.-B. Song, and W. Chung, "Mobile Robot Localization Using Fusion of Object Recognition and Range Information," in *Proceedings of IEEE International Conference on Robotics and Automation (ICRA)*, (Roma, Italy), pp. 3533–3538, IEEE, 2007.

[2] 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 *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, (Boston, USA), pp. 1352–1359, IEEE, 2013.

[3] M. Crocco, A. D. Bue, I. B. Barbosa, and V. Murino, "A Closed Form Solution for the Self-Calibration of Heterogeneous Sensors," in *Procedings of the British Machine Vision Conference 2012*, (Los Angeles, California), 2012.

[4] M. Crocco, A. D. Bue, and V. Murino, "A Bilinear Approach to the Position Self-Calibration of Multiple Sensors," *IEEE Transactions on Signal Processing*, vol. 60, pp. 660–673, Feb. 2012.

[5] P. Biswas, H. Aghajan, and Y. Ye, "Integration of Angle of Arrival Information for Multimodal Sensor Network Localization using Semidefinite Programming," in *Proceedings of 39th Asilomar Conference on Signals, Systems and Computers*, 2005.

[6] A. Beck and P. Stoica, "Exact and Approximate Solutions of Source Localization Problems," *IEEE Transactions on Signal Processing*, vol. 56, pp. 1770–1778, May 2008.

[7] P. Oguz-Ekim, J. Gomes, J. Xavier, M. Stosic, and P. Oliveira, "An Angular Approach for Range-Based Approximate Maximum Likelihood Source Localization Through Convex Relaxation," *IEEE Transactions on Wireless Communications*, vol. 13, pp. 3951–3964, July 2014.

[8] S. Garrido-Jurado, R. Muñoz-Salinas, F. J. Madrid-Cuevas, and M. J. Marín-Jiménez, "Automatic generation and detection of highly reliable fiducial markers under occlusion," *Pattern Recognition*, vol. 47, no. 6, pp. 2280 – 2292, 2014.

[9] N. B. Priyantha, A. Chakraborty, and H. Balakrishnan, "The Cricket Location-Support System," in *The Proceedings of the sixth ACM International Conference on Mobile Computing and Networking (ACM Mobicom 2000)*, pp. 32–43, 2000.