A Closed Form Solution for the Self-Calibration of Heterogeneous Sensors

Marco Crocco
marco.crocco@iit.it
Alessio Del Bue
alessio.delbue@iit.it
Igor Barros Barbosa
igorbb@gmail.com
Vittorio Murino
vittorio.murino@iit.it

Pattern Analysis & Computer Vision - PAVIS
Istituto Italiano di Tecnologia - IIT
Via Morego, 30, 16163 Genova, Italy

Abstract

We present a novel closed-form solution for the joint self-calibration of video and range sensors. The approach single assumption is the availability of synchronous time of flight (i.e., range distances) measurements and visual position of the target on images acquired by a set of cameras. In such case, we make explicit a rank constraint that is valid for both image and range data. This rank property is used to find an initial and affine solution via bilinear factorization, which is then corrected by enforcing the metric constraints characteristic for both sensor modalities (i.e., camera and anchors constraints). The output of the algorithm is the identification of the target/range sensor position and the calibration of the cameras. The application extent of our approach is broad and versatile. In fact, with the same framework, we can deal with, but not restricted to, two very different applications. The first is aimed at calibrating cameras and microphones deployed in an unknown environment. The second uses a RGB-D device to reconstruct the 3D position of a set of keypoints using the camera and depth map images. Synthetic and real tests show the algorithm performance under different levels of noise and configurations of target locations, number of sensors and cameras.

1 Introduction

The technological advancement of distributed sensor architectures generates systems that are inherently multi-modal. In particular, different sensor modalities are often coupled together in order to compensate for the deficiencies of a single-modality alone. For instance, video cameras are certainly the most versatile sensors that might reach very high-resolutions at high frame rates while being subject to all the problems given by saturation, lens aberrations, etc. Such problems might affect Computer Vision algorithms dedicated to a specific task. As an example, if we have to detect the background and foreground objects, the variability of the object texture can render the problem rather complex. Coupling a range device could clearly simplify the problem since segmentation in depth images, in comparison with the vision counterpart, might be considered trivial. Obviously, a depth device alone would not preserve the appearance of objects and might be restricted to lower resolution, so that a coupled

http://dx.doi.org/10.5244/C.26.25

camera is an essential aid for tasks such as object recognition. Similarly, for audio data, vision has been coupled with the output of a microphone in order to robustify the localisation of a given target [1, 4, 13]. These methods normally require an accurate calibration of both cameras and microphones, a solution in most cases impractical if not using a particular and a priori known displacement of the sensors. These are just a few examples, however there exists multi-modal algorithms that use inputs of video together with thermic images [16], RFID [7, 11], etc. to obtain a more robust task-driven perfomance.

In order to obtain a successful sensor data fusion, a key issue is to obtain a precise alignment of the different sensors. In particular, we deal with the case of video and range sensors where for the latter we consider any devices that can measure a distance from a target. Moreover, we specifically deal with the case of several sensors deployed in an unknown area and unknown positions of targets. This scenario for instance fits well in the case of heterogeneous sensor networks but also for single devices such as the Kinect that can measure several distances (i.e., a depth image) taken in an indoor area. Even if the setups are different, all these problems result in the same *self-calibration* problem which aims to the simultaneous localisation of sensors and targets.

In particular, this problem is rather critical when attempting to localise a certain target immersed in a sensor network. Such problem is possibly the most interesting and, often, the same very reason why a network is deployed. As an example, in wireless sensor networks one would like to monitor in real-time the location of devices of interests [17] such as the equipments in a hospital. In video-surveillance, dangerous targets should be localised before the situation may degenerate [5]. Military applications might also use a sensor network in order to identify ground targets [8] or hidden snipers in a combat area [24]. Workers may also be localised in order to warn them that they are trespassing a dangerous area in a factory [9].The common aspect among all these scenarios is that they all need precise measurements of the target position. Indeed, this is possible only if the calibration of the heterogeneous sensors is known and accurate. Though in recent years some approaches [10, 12] have been proposed for the joint calibration of range sensors and video cameras, they require a specific calibration pattern and, as a consequence, cannot in general be easily adapted to the aforementioned problem.

In this paper, we introduce a novel self-calibration problem where cameras and range sensors are measuring the position of a target. In such problem, both the positions of sensors, cameras and the target are unknown apart from the knowledge of (few) sensor anchors. Interestingly, if the anchors are not available, the system is still able to provide a solution using solely metric constraints from the cameras. The solution of the self-calibration problem will provide the 3D localisation of the target, the position of the range sensors and the cameras calibration.

The remainder of the paper introduces the formalisation of the self-calibration problem in the range (Sec. 2) and camera (Sec. 3) case. The following Sec. 4 defines the joint self-calibration as a factorization problem with specific metric constraints given by both modalities. Sec. 5 presents the synthetic results for a typical multi-view camera and microphone calibration problem while the real test uses data coming from a Kinect device. Conclusions in Sec. 6 draw the path for possible extensions of the proposed method.

2 Range sensors calibration

Let us consider m range sensors and n targets that generate the range signal detected by the sensors, both of them laying in a 3D space. The sensor and target coordinates can be stored

respectively in a $m \times 3$ matrix $\mathtt{S}$ and an $3 \times n$ matrix $\mathtt{T}$ defined as follows:

$$\mathtt{S} = \begin{bmatrix} s_{11} & s_{12} & s_{13} \\ s_{21} & s_{22} & s_{23} \\ \vdots & \vdots & \vdots \\ s_{m1} & s_{m2} & s_{m3} \end{bmatrix} \quad \text{and} \quad \mathtt{T} = \begin{bmatrix} t_{11} & t_{21} & \cdots & t_{n1} \\ t_{12} & t_{22} & \cdots & t_{n2} \\ t_{13} & t_{23} & \cdots & t_{n3} \end{bmatrix}. \tag{1}$$

where s_{il} and t_{il} are the l-th coordinate of the i-th range sensor and target respectively. The first a sensors in $\mathtt{S}$, with $a \geq 0$, are considered to be anchors i.e. their 3D coordinates are *a priori* known, whereas the coordinates of remaining $m-a$ sensors and the coordinates of all targets are unknown. Each range sensor, including the anchors, is supposed to estimate a distance between itself and each of the targets. For the sake of generality we do not consider a specific procedure for the distance estimation but we can mention that the great part of range sensors relies on two principles. First, the time of flight between sensor and target is measured and distance is founded simply dividing by the signal velocity in the medium. Second, the distance is estimated by measuring the signal intensity at the sensor and comparing it with the known emission strength. The square of the estimated distances d_{ij} between generic sensor i and target j can be stored in an $m \times n$ matrix $\mathtt{D}$ defined as follows:

$$\mathtt{D} = \begin{bmatrix} d_{11}^2 & d_{12}^2 & \cdots & d_{1n}^2 \\ d_{21}^2 & d_{22}^2 & \cdots & d_{2n}^2 \\ \vdots & \vdots & \ddots & \vdots \\ d_{m1}^2 & d_{m2}^2 & \cdots & d_{mn}^2 \end{bmatrix}. \tag{2}$$

In an ideal situation, assuming that each estimated distance is equal to the actual one, the following set of nm equations hold for $i = 1 \dots m$ and $j = 1 \dots n$:

$$s_{i1}^2 + s_{i2}^2 + s_{i3}^2 + t_{j1}^2 + t_{j2}^2 + t_{j3}^2 - 2s_{i1}t_{j1} - 2s_{i2}t_{j2} - 2s_{i3}t_{j3} = d_{i,j}^2. \tag{3}$$

In order to obtain a bilinear form in the sensors and events coordinate vectors, the first six quadratic terms in the above equations have to be eliminated [6]. To this aim one can subtract the $1,j$-th equation to the i,j-th equation in (3) for $i = 2 \dots m$ and $j = 1 \dots n$, obtaining a set of $(m-1)n$ equations.

$$\begin{aligned} &s_{i1}^2 + s_{i2}^2 + s_{i3}^2 - (s_{11}^2 + s_{12}^2 + s_{13}^2) - 2(s_{i1} - s_{11})t_{j1} \\ &- 2(s_{i2} - s_{12})t_{j2} - 2(s_{i3} - s_{13})t_{j3} = d_{i,j}^2 - d_{1,j}^2. \end{aligned} \tag{4}$$

In the same way, subtracting the $i,1$-st equation to the i,j-th equation in (4) for $i = 2 \dots m$ and $j = 2 \dots n$, one obtains a set of $(m-1)(n-1)$ equations:

$$\begin{aligned} &-2(s_{i1} - s_{11})(t_{j1} - t_{11}) - 2(s_{i2} - s_{12})(m_{j2} - m_{12}) + \\ &-2(s_{i3} - s_{13})(t_{j3} - t_{13}) = d_{i,j}^2 - d_{1,j}^2 - d_{i,1}^2 + d_{1,1}^2. \end{aligned} \tag{5}$$

The very same operations can be expressed in matrix form by first defining the following special vectors/matrix:

$$\mathbf{e}_j^\top = (0, \dots, 0, 1, 0, \dots, 0), \quad \mathbf{1}_b^\top = (1, \dots, 1), \quad \mathtt{P}_b = \begin{bmatrix} \mathbf{0} & \mathtt{I} \end{bmatrix}.$$

where $\mathbf{e}_j$ represents a vector of zeros with a single 1 at position j and $\mathbf{1}_b$ represents a vector of b ones. The square matrices $\mathtt{P}_{(b-1)\times b}$ and $\mathtt{P}_{b\times(b-1)}^\top$ instead remove the first row and column

via left and right matrix multiplications respectively. It is possible to remove the quadratic terms from $\mathtt{D}$ such as:

$$\hat{\mathtt{D}} = \mathtt{D} - \mathbf{1}_m \mathbf{e}_1^\top \mathtt{D} - (\mathtt{D} - \mathbf{1}_m \mathbf{e}_1^\top \mathtt{D})\, \mathbf{e}_1 \mathbf{1}_n^\top = \mathtt{D} - \mathbf{1}_m \mathbf{e}_1^\top \mathtt{D} - \mathtt{D}\,\mathbf{e}_1 \mathbf{1}_n^\top + \mathbf{1}_m \mathbf{e}_1^\top \mathtt{D}\,\mathbf{e}_1 \mathbf{1}_n^\top. \tag{6}$$

Now we can eliminate the zero-row and zero-column of $\hat{\mathtt{D}}$ such as:

$$\tilde{\mathtt{D}} = \mathtt{P}_m \hat{\mathtt{D}} \mathtt{P}_n^\top = \begin{bmatrix} \tilde{d}_{1,1} & \tilde{d}_{1,2} & \cdots & \tilde{d}_{1,n-1} \\ \tilde{d}_{2,1} & \tilde{d}_{2,2} & \cdots & \tilde{d}_{2,n-1} \\ \vdots & \vdots & \ddots & \vdots \\ \tilde{d}_{m-1,1} & \tilde{d}_{m-1,2} & \cdots & \tilde{d}_{m-1,n-1} \end{bmatrix}. \tag{7}$$

where $\tilde{d}_{i-1,j-1} = d_{i,j}^2 - d_{1,j}^2 - d_{i,1}^2 + d_{1,1}^2$. Let us organize the remaining terms in (5) in the following matrices as:

$$\tilde{\mathtt{S}} = \begin{bmatrix} s_{21} - s_{11} & s_{22} - s_{12} & s_{23} - s_{13} \\ s_{31} - s_{11} & s_{32} - s_{12} & s_{33} - s_{13} \\ \vdots & \vdots & \vdots \\ s_{m1} - s_{11} & s_{m2} - s_{12} & s_{m3} - s_{13} \end{bmatrix}, \quad \tilde{\mathtt{T}} = \begin{bmatrix} t_{21} - t_{11} & t_{31} - t_{11} & \cdots & t_{n1} - t_{11} \\ t_{22} - t_{12} & t_{32} - t_{12} & \cdots & t_{n2} - t_{12} \\ t_{23} - t_{13} & t_{33} - t_{13} & \cdots & t_{n3} - t_{13} \end{bmatrix}; \tag{8}$$

where $\tilde{\mathtt{S}} = \mathtt{P}_m\,(\mathtt{S} - \mathbf{1}_m \mathbf{e}_1^\top \mathtt{S})$ and $\tilde{\mathtt{T}} = (\mathtt{T} - \mathtt{T}\mathbf{e}_1 \mathbf{1}_n^\top)\,\mathtt{P}_n$. Using the previously defined matrices the set of equations in (5) can be expressed in a matrix form as:

$$-2\tilde{\mathtt{S}}\tilde{\mathtt{T}} = \tilde{\mathtt{D}}. \tag{9}$$

The $(m-1) \times (n-1)$ matrix $\tilde{\mathtt{D}}$ has rank equal to three since it is a product between the $(m-1) \times 3$ matrix $-2\tilde{\mathtt{S}}$ and the $3 \times (n-1)$ matrix $\tilde{\mathtt{T}}$. If we apply a SVD to the data matrix $\tilde{\mathtt{D}}$ we have, in case of no noise, that the singular values after the third are equal to zero. Thus we can truncate these SVD components such as:

$$\mathtt{U}\mathtt{V}\mathtt{W} = \tilde{\mathtt{D}}, \tag{10}$$

where $\mathtt{U}$ is an $(m-1) \times 3$ matrix, $\mathtt{V}$ is a 3×3 diagonal matrix and $\mathtt{W}$ is a $3 \times (n-1)$ matrix. In a practical situation, in presence of measurement noise, the rank of $\tilde{\mathtt{D}}$ will be probably higher than three: in this case only the three biggest singular values in $\mathtt{V}$ will be considered reducing the size of $\mathtt{U}$, $\mathtt{V}$ and $\mathtt{W}$ according to the noise-free case. From (9) and (10), for every invertible 3×3 matrix $\mathtt{C}$, the following relationships hold:

$$-2\tilde{\mathtt{S}} = \mathtt{U}\mathtt{Q}_s \quad \text{and} \quad \tilde{\mathtt{T}} = \mathtt{Q}_s^{-1}\mathtt{V}\mathtt{W}.$$

The matrix $\mathtt{Q}_s$ is called the "mixing matrix" since it mixes the components obtained from the SVD in order to obtain the correct solution given the original sensors localization problem. The matrix $\mathtt{Q}_s$ can be found exploiting the a priori knowledge on anchors position by the following procedure. By defining the matrix $\bar{\mathtt{A}}$ as the first $a-1$ rows of matrix $\mathtt{U}$, and the matrix $\tilde{\mathtt{A}}$ of dimension $(a-1) \times 3$ as

$$\tilde{\mathtt{A}} = \mathtt{P}_a\,(\mathtt{A} - \mathbf{1}_a \mathbf{e}_1^\top \mathtt{A}) \tag{11}$$

where $\mathtt{A}$ is the matrix of the known coordinates of anchors, the following equality holds:

$$\bar{\mathtt{A}}\mathtt{Q}_s = -2\tilde{\mathtt{A}} \tag{12}$$

which can be solved with LS if at least four anchors are present. If no anchors are available the relative position of sensors and targets can be found for less than a 3D translation and rotation. In such case, the mixing matrix $\mathtt{Q}_s$ can be derived in closed form [6] providing that at least one sensor is coincident with one target. Alternatively, if at least five targets and ten sensors or *viceversa* are available, the closed form solution proposed in [22] can be adopted.

3 Video cameras calibration

In this work, we define the camera self-calibration problem as the contemporaneous computation of the parameters of an affine camera and the 3D position of a target solely from image measurements. In particular, a target can be defined as a collection of a set of 3D points stored in a matrix $\mathtt{T}$ as defined in the previous section. Each of the n points is projected onto each camera image frame using a camera projection matrix such that:

$$\begin{pmatrix} u_{kj} \\ v_{kj} \end{pmatrix} = \left[\begin{array}{c|c} \mathtt{R}_k & \mathbf{z}_k \end{array}\right] \begin{pmatrix} t_{j1} \\ t_{j2} \\ t_{j3} \\ 1 \end{pmatrix} \tag{13}$$

where u_{kj} and v_{kj} represents the two image coordinates of the target j as seen by camera k. The 2×3 matrix $\mathtt{R}_k$ and the 2-vector $\mathbf{z}_k$ are the parameters of the camera that is considered to be affine in such scenario. The affine camera matrix approximation holds in real scenarios where the target is imaged at distance so favoring in particular outdoor sensor networks. Given the set of targets $\mathtt{T}$ we can write the image coordinates of the target as:

$$\mathtt{G}_k = \begin{pmatrix} u_{k1} & \cdots & u_{kn} \\ v_{k1} & \cdots & v_{kn} \end{pmatrix} = \left[\begin{array}{c|c} \mathtt{R}_k & \mathbf{z}_k \end{array}\right] \begin{pmatrix} t_{11} & \cdots & t_{n1} \\ t_{12} & \cdots & t_{n2} \\ t_{13} & \cdots & t_{n3} \\ 1 & \cdots & 1 \end{pmatrix} = \left[\begin{array}{c|c} \mathtt{R}_k & \mathbf{z}_k \end{array}\right] \begin{bmatrix} \mathtt{T} \\ \mathbf{1}^\top \end{bmatrix}, \tag{14}$$

where the matrix $\mathtt{G}_k$ of size $2 \times n$ contains the image target positions for the camera k. With multiple cameras and a single target moving, we can build a set of equations such as:

$$\mathtt{G} = \begin{bmatrix} \mathtt{G}_1 \\ \vdots \\ \mathtt{G}_c \end{bmatrix} = \begin{bmatrix} \left[\begin{array}{c|c} \mathtt{R}_1 & \mathbf{z}_1 \end{array}\right] \\ \vdots \\ \left[\begin{array}{c|c} \mathtt{R}_c & \mathbf{z}_c \end{array}\right] \end{bmatrix} \begin{bmatrix} \mathtt{T} \\ \mathbf{1}^\top \end{bmatrix} = \mathtt{C} \begin{bmatrix} \mathtt{T} \\ \mathbf{1}^\top \end{bmatrix}, \tag{15}$$

where the $2c \times 4$ matrix $\mathtt{C}$ contains the c camera matrices. The bilinear form of equation (15) implies a rank 4 constraint on the image measurement matrix $\mathtt{G}$. In the case of affine cameras, the image measurements may be registered to the image centroid of the cameras or aligned to a selected image point (e.g. the first target in $\mathtt{T}$) such as:

$$\tilde{\mathtt{G}} = (\mathtt{G} - \mathbf{g}_1 \mathbf{1}_n^\top)\mathtt{P}_n^\top = \begin{bmatrix} \mathtt{R}_1 \\ \vdots \\ \mathtt{R}_c \end{bmatrix} \tilde{\mathtt{T}} = \tilde{\mathtt{C}}\tilde{\mathtt{T}} \tag{16}$$

where $\mathbf{1}_n$ is a vector of n ones and $\mathbf{g}_1$ is a $2c \times 1$ vector containing the image measurements of the first target i.e. $\mathbf{g}_1 = [u_{11}, v_{11}, \ldots, u_{c1}, v_{c1}]^\top$. The matrix form in eq. (16) is a classical

formulation of the Structure from Motion (SfM) problem for a moving camera. Such formulation is popular in Computer Vision because it led to efficient closed form solutions to both 3D reconstruction and camera calibration. The seminal work of Tomasi and Kanade [27] first proposed an efficient solution for the orthographic camera case that was successfully extended to more complex camera models (e.g. [25], [15] to cite a few).

To summarise, given the input data in $\tilde{\mathtt{G}}$, it is possible to obtain a first decomposition of the bilinear components by means of a standard SVD truncated to rank 3 in a very similar manner as obtained for the distance measurements in eq. (10). With an abuse of notation we keep the same symbols for the SVD values giving $\tilde{\mathtt{G}} = \mathtt{U}\mathtt{V}\mathtt{W}$ where $\mathtt{U}$ is a $2c \times 3$ matrix, $\mathtt{V}$ is a 3×3 diagonal matrix and $\mathtt{W}$ is a $3 \times (n-1)$ matrix. Similarly as in the range sensor case, this decomposition is not unique and there exists a mixing matrix $\mathtt{Q}_v$ such that:

$$\tilde{\mathtt{C}} = \mathtt{U}\mathtt{Q}_v \quad \text{and} \quad \tilde{\mathtt{T}} = \mathtt{Q}_v^{-1}\mathtt{V}\mathtt{W} \tag{17}$$

The key element for solving $\mathtt{Q}_v$ is to exploit the constraints imposed by the specific camera models. Notice that the affine camera matrix $\mathtt{R}_k$ can assume specific forms given such constraints; in particular we have that is a scaled orthographic camera matrix if $\mathtt{R}_k\mathtt{R}_k^\top = s_k\mathtt{I}_2$. Now notice that each camera model enforces a specific constraint on the elements of $\tilde{\mathtt{C}}$. Thus we compute the matrix $\mathtt{Q}_v$ that enforces exactly the constraint in the chosen camera model. In order to simplify the notation, from this point we will choose the scaled orthographic camera model, however notice that the proposed solution might be applied for more descriptive camera models [14, 15, 21]. Thus, if we divide the $\mathtt{U}$ matrix from SVD into 2×3 sub-blocks as $\mathtt{U}^\top = \begin{bmatrix}\mathtt{U}_1^\top & \cdots & \mathtt{U}_c^\top\end{bmatrix}$ we have that $\mathtt{Q}_v$ has to satisfy a set of c equations such that:

$$\mathtt{U}_k\mathtt{Q}_v\mathtt{Q}_v^\top\mathtt{U}_k^\top = s_k\mathtt{I}_2. \tag{18}$$

These constraints are quadratic in $\mathtt{Q}_v$ therefore not solvable in closed form. However, a convenient solution is found by considering the symmetric matrix $\mathtt{H}_v = \mathtt{Q}_v\mathtt{Q}_v^\top$ and by rewriting the system in the 6 unknowns of $\mathtt{H}_v$. For a scaled orthographic camera matrix we have two equations for each camera thus requiring at least 3 cameras for obtaining a unique solution for $\mathtt{H}_v$. After recovering $\mathtt{H}_v$ via LS, the matrix $\mathtt{Q}_v$ can be found with a Cholesky decomposition [1]. The LS equations for the matrix $\mathtt{H}_v$ will be presented in the next section since they will lead to the joint solution of the self-calibration problem.

4 Joint self-calibration of visual and range sensors

The common property for solving jointly the self-calibration problem is that both measured data sussist on a common subspace as defined by the target position $\mathtt{T}$. This fact emerges clearly only after reducing both the measured data to rank 3 bilinear models in the respective domains. The consequence is that the fusion of the modalities is for the first time strictly geometrical, in the sense that the data is now explicitly linked by the metric position of the targets. This leads to the possibility of computing a joint closed form solution using the range-visual constraints of the heterogeneous sensors. The measurements from range and visual sensors form the measurement matrix $\mathtt{Y}$ of size $(m+2c-1) \times (n-1)$ such as:

$$\mathtt{Y} = \begin{bmatrix}\tilde{\mathtt{D}} \\ \tilde{\mathtt{G}}\end{bmatrix} = \begin{bmatrix}-2\tilde{\mathtt{S}} \\ \tilde{\mathtt{C}}\end{bmatrix}\tilde{\mathtt{T}}.$$

[1] Choleski decomposition requires the matrix $\mathtt{H}_v$ to be positive definite. This requirement is satisfied most of the time with mild noise conditions and without outliers in the measured data. If $\mathtt{H}_v$ is not positive definite after the least square solution, it is possible to revert to a SDP problem where the condition $\mathtt{H}_v \succ 0$ can be imposed explicitly.

Here the same target position has to be subtracted from both range and image measurements. Thus a single SVD can be used in order to obtain a first factorization that can be then upgraded to metric with a custom solution. Notice that it might be necessary a data normalization in order to balance the results of SVD. This is because the range of values between image coordinates might be rather different [2]. As a rule of the thumb, we compute a scale factor such that $\|\tilde{\mathtt{D}}\|_F = \|\tilde{\mathtt{G}}\|_F$ so as to evaluate equally visual and range data when performing SVD. Now, the SVD provides again a tern $\mathtt{UVW}$ that has to be transformed by a 3×3 mixing matrix $\mathtt{Q}_j$ such that:

$$\mathtt{U}\mathtt{Q}_j = \begin{bmatrix}\mathtt{U}_s\\ \mathtt{U}_v\end{bmatrix}\mathtt{Q}_j = \begin{bmatrix}-2\tilde{\mathtt{S}}\\ \tilde{\mathtt{C}}\end{bmatrix}. \tag{19}$$

The mixing matrix $\mathtt{Q}_j$ can be found by putting together the constraints related to both anchor positions and scaled orthographic camera models. In particular for the image data, constraints related to the k-th camera are given by $\mathtt{U}_k\mathtt{H}\mathtt{U}_k^\top = s_k\mathtt{I}_2$ where $\mathtt{H}$ is a symmetric matrix defined as $\mathtt{H} = \mathtt{Q}_j\mathtt{Q}_j^\top$, as in Eq. (18). The constraints related to the i-th camera can be recast in two equations giving [15]:

$$\mathbf{u}_{i1}^\top\mathtt{H}\mathbf{u}_{i2} = 0 \;\text{ and }\; \mathbf{u}_{i1}^\top\mathtt{H}\mathbf{u}_{i1} - \mathbf{u}_{i2}^\top\mathtt{H}\mathbf{u}_{i2} = (\mathbf{u}_{i1}-\mathbf{u}_{i2})^\top\mathtt{H}(\mathbf{u}_{i1}+\mathbf{u}_{i2}) = 0, \tag{20}$$

where $\mathbf{u}_{i1}$ and $\mathbf{u}_{i2}$ constitute the i-th block of the matrix $\mathtt{U}_v$ as follows:

$$\mathtt{U}_i = \begin{bmatrix}\mathbf{u}_{i1}^\top\\ \mathbf{u}_{i2}^\top\end{bmatrix}. \tag{21}$$

These equations in (20) can be written in vector form [3] for the elements of $\mathtt{H}$ defining:

$$vc(\mathbf{u}_{i1},\mathbf{u}_{i2}) = vech(\mathbf{u}_{i1}\mathbf{u}_{i2}^\top - \mathbf{u}_{i2}\mathbf{u}_{i1}^\top - diag(\mathbf{u}_{i1}\odot\mathbf{u}_{i2})), \tag{22}$$

where $vech()$ is the operator that vectorizes symmetric matrices. Thus, we can define a LS problem for each camera using (22), such that:

$$\hat{\mathtt{U}}_i vech(\mathtt{H}) = 0 \;\text{ where }\; \hat{\mathtt{U}}_i = \begin{bmatrix}(vc((\mathbf{u}_{i1},\mathbf{u}_{i2})))^\top\\ (vc(\mathbf{u}_{i1}-\mathbf{u}_{i2},\mathbf{u}_{i1}+\mathbf{u}_{i2}))^\top\end{bmatrix}. \tag{23}$$

Similarly, we have now to embed the range constraints into a LS form in order to solve jointly for the video-range constraints. If we consider again Eq. (12), we have that:

$$\bar{\mathtt{A}}\mathtt{H}\bar{\mathtt{A}}^\top = 4\tilde{\mathtt{A}}\tilde{\mathtt{A}}^\top. \tag{24}$$

The matrix equation (24) can be written element-wise as follows:

$$\bar{\mathbf{a}}_i^\top\mathtt{H}\bar{\mathbf{a}}_j = \tilde{\mathbf{a}}_i^\top\tilde{\mathbf{a}}_j, \tag{25}$$

for $i,j = 1\ldots a-1$ where $\bar{\mathbf{a}}_i^\top$ and $\tilde{\mathbf{a}}_i^\top$ are the i-th row of matrices $\bar{\mathtt{A}}$ and $4\tilde{\mathtt{A}}$ respectively. Exploiting the same formalism defined in (22), Eq. (25) can be written as:

$$\hat{\mathbf{a}}_{ij} vech(\mathtt{H}) = \tilde{\mathbf{a}}_i^\top\tilde{\mathbf{a}}_j \quad \text{where} \quad \hat{\mathbf{a}}_{ij} = vc(\bar{\mathbf{a}}_i,\bar{\mathbf{a}}_j). \tag{26}$$

[2] For instance, current HD cameras allows the image coordinates in $\mathtt{G}$ to range from zero to 1920 while the distance values measured depends by the chosen metric unit.

Since $vc()$ is a symmetric operator, only $(a-1)a/2$ of $(a-1)(a-1)$ linear constraints defined in (26) are independent. As a consequence, the anchor constraints can be expressed in a matrix form as $\hat{\mathtt{A}}\, vech(\mathtt{H}) = \check{\mathbf{a}}$, where the $(a-1)a/2 \times 3$ matrix $\hat{\mathtt{A}}$ and the $(a-1)a/2$ vector $\check{\mathbf{a}}$ are defined as:

$$\hat{\mathtt{A}}^\top = \begin{bmatrix} \hat{\mathbf{a}}_{11}^\top & \cdots & \hat{\mathbf{a}}_{1a-1}^\top & \hat{\mathbf{a}}_{22}^\top & \cdots & \hat{\mathbf{a}}_{a-1a-1}^\top \end{bmatrix} \tag{27}$$

$$\check{\mathbf{a}}^\top = \begin{pmatrix} \tilde{\mathbf{a}}_1^\top \tilde{\mathbf{a}}_1 & \cdots & \tilde{\mathbf{a}}_1^\top \tilde{\mathbf{a}}_{a-1} & \tilde{\mathbf{a}}_2^\top \tilde{\mathbf{a}}_2 & \cdots & \tilde{\mathbf{a}}_{a-1}^\top \tilde{\mathbf{a}}_{a-1} \end{pmatrix}. \tag{28}$$

This last step allows to write the final equation embodying the camera and range sensor constraints:

$$\begin{bmatrix} \hat{\mathtt{A}} \\ \hat{\mathtt{U}} \end{bmatrix} vech(\mathtt{H}) = \begin{bmatrix} \check{\mathbf{a}} \\ \mathbf{0} \end{bmatrix}, \quad \text{where} \quad \hat{\mathtt{U}}^\top = \begin{bmatrix} \hat{\mathtt{U}}_1^\top & \hat{\mathtt{U}}_2^\top & \cdots & \hat{\mathtt{U}}_c^\top \end{bmatrix}. \tag{29}$$

Equation (29) can be solved with a pseudo-inverse if at least 6 equations are present:

$$vech(\mathtt{H}) = (\mathtt{F}^\top \mathtt{F})^{-1} \mathtt{F}^\top \mathbf{r} \quad \text{where} \quad \mathtt{F} = \begin{bmatrix} \hat{\mathtt{A}} \\ \hat{\mathtt{U}} \end{bmatrix} \quad \text{and} \quad \mathbf{r} = \begin{bmatrix} \check{\mathbf{a}} \\ \mathbf{0} \end{bmatrix}. \tag{30}$$

Notice here that, differently from the single modality case, the minimal number of equations can be mixed between cameras and range sensor (i.e. we do not need a minimum of 3 cameras and 4 anchors in order to have a solvable problem). In general each camera provide two constraints, while a anchors provide $(a-1)a/2$ constraints, as can be seen from the previous equations. As a consequence, an interesting mixed minimal configurations exist of two cameras and three anchors, besides the configuration with no anchors (three cameras) and no cameras (four anchors). Finally, notice that if four anchors are available, the proposed approach has the further property of eliminating the rotational *gauge* freedom of the problem [2, 13].

5 Experiments

Synthetic Test. A set of synthetic experiments has been performed in order to evaluate the self-calibration algorithm with varying numbers of cameras, range sensors and targets. This setup is normally representative of a sensor calibration scenario [8, 26] and target localisation problems from randomly displaced sensors [19, 23]. Cameras, sensors and targets were randomly displaced into a cubic region according to a uniform distribution. A measurement noise was added to the distances measured by range sensors and to the pixel position of targets in camera images. We defined the noise magnitude in relation to the data Frobenius norm in order to make them comparable. Thus we introduced two normalized noise measures N_r (range error) and N_c (camera error) such that:

$$N_r = \|\mathtt{N}_1\|_F / \|\mathtt{D}\|_F \quad \text{and} \quad N_c = \|\mathtt{N}_2\|_F / \|\mathtt{W}\|_F, \tag{31}$$

where $\|\cdot\|_F$ is the Frobenius norm, $\mathtt{D}$ is the matrix of distances among range sensors and targets, $\mathtt{N}_1$ and $\mathtt{N}_2$ are two matrices of size $m \times n$ and $2c \times n$ respectively whose elements are realizations of a zero mean Gaussian PDF. Similarly, in order to quantify the error on the position estimation of targets the following metric is adopted:

$$Et = \|\mathtt{T} - \mathtt{T}_{GT}\|_F / \|\mathtt{T}_{GT}\|_F, \tag{32}$$

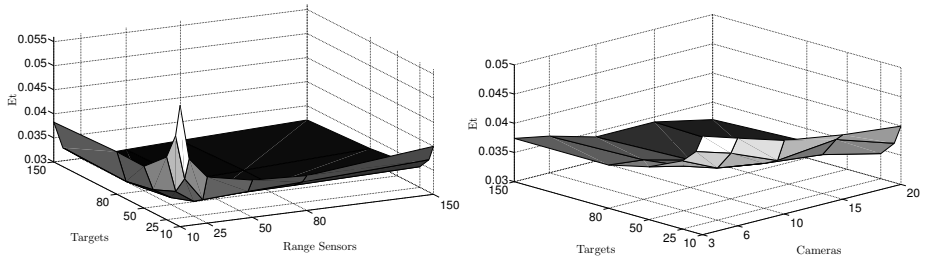

Figure 1: Left image shows a surface plot of the target reconstruction 3D error with respect to varying number of targets and cameras. The right image shows the same evaluation with respect to varying number of targets and range sensors.

where $\mathtt{T}_{GT}$ is the ground truth matrix of targets 3D coordinates whereas $\mathtt{T}$ is its estimated counterpart. Each experiment has been repeated 2000 times, each time changing the random 3D configuration of cameras, range sensors and targets and the noise. The left plot of Figure 1 shows the average Et in function of the number of targets $(10, 15, 25, 50, 80, 150)$, range sensors $(10, 15, 25, 50, 80, 150)$, 20 cameras, 5 anchors, $N_r = 0.028$ and $N_c = 0.013$ (we simulated the likely case when range measurements are less precise than image measurements). At a first sight, it is possible to see that the error surface is decreasing at the increase of the number of targets/sensors reaching a minimum value of $Et = 0.0301$ for the target/sensor pair $(150, 25)$. The maximum error of $Et = 0.055$ is achieved for the pair $(10, 10)$. In general, a quasi-constant value of Et (about 0.032) is observed for above 25 sensors and targets.exceeeds 50. This last behaviour is easily explanaible considering that, as $N_r > N_c$, increasing the number of range sensors in respect to cameras the more noisy range data prevail in the measurement matrix $\mathtt{Y}$. The right plot of Figure 1 shows the average Et in function of the number of targets $(10, 15, 25, 50, 80, 150)$ and cameras $(3, 6, 10, 15, 20)$, 150 range sensors and keeping unaltered the other parameters. In this case the highest reconstruction error ($Et = 0.047$) is obtained for 10 targets and 3 cameras, whereas the lowest one ($Et = 0.032$) is obtained for 150 targets and 20 cameras. Notice that Et is monotonically decreasing with the number of cameras except for 15 range sensors where a slight increase is present moving from 10 to 15 cameras.

Real Test. The new brand of RGB-D devices generates a high number of range measurements coupled with standard images. Here, our targets are represented by a set of image point trajectories tracked in a video grabbed by a Kinect sensor (check Figure 2a). Depth and standard images were previously aligned using the proprietary intrinsic camera model of PrimeSense by using their development framework *OpenNI* [20]. In such way, at each image coordinate would correspond the correct depth value. By applying our self-calibration algorithm to the data matrix $\mathtt{Y}$ of size 533×69 we obtained the 3D position of each point (the targets) together with the 3D location of the depth sensor and standard camera parameters. The several planar surfaces of the scene have been reconstructued correctly given both information from depth and standard images.

6 Conclusions

To the authors knowledge, for the first time we have presented a new geometrical constraint for the fusion of information acquired from standard video camera and range sensors. Such theoretical result leads to a closed-form solution for the self-calibration of heterogeneous sensors that is able to accurately localise the 3D position of both sensors and the targets. This

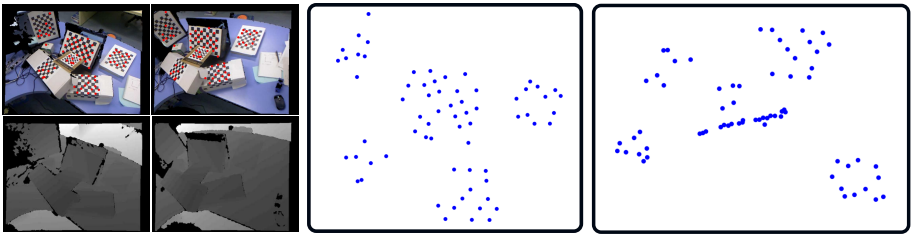

Figure 2: The Kinect test. a) shows on top two image samples of the image sequence used for the joint reconstruction. The figures on the bottom are the corresponding depth images. b) shows a frontal view of the 3D reconstruction while c) presents a side view aligned with the biggest box in the scenario showing the expected planarity of the object.

new result is of practical application in many scenarios. Future works will be dedicated to the use of this new constraint in several audio-video applications such as speaker localisation and tracking. On the theoretical aspects, an extension to the missing data case is a direction to follow in order to deal with the most complex network topologies.

References

[1] Xavier Alameda-Pineda, Vasil Khalidov, Radu Horaud, and Florence Forbes. Finding audio-visual events in informal social gatherings. In *Proceedings of the 13th international conference on multimodal interfaces*, ICMI '11, pages 247–254, New York, NY, USA, 2011. ACM. ISBN 978-1-4503-0641-6. doi: 10.1145/2070481.2070527.

[2] A. Bartoli. Towards gauge invariant bundle adjustment: A solution based on gauge dependent damping. In *Ninth IEEE International Conference on Computer Vision, 2003*, pages 760–765. IEEE, 2003.

[3] M. Brand. A direct method for 3d factorization of nonrigid motion observed in 2d. In *IEEE conference on Computer Vision and Pattern Recognition (CVPR 2005)*, volume 2, pages 122 – 128, June 2005.

[4] A. Brutti and O. Lanz. A joint particle filter to track the position and head orientation of people using audio visual cues. In *Proc. of EUSIPCO*, 2010.

[5] Chung-Kuo Chang and J. Huang. Video surveillance for hazardous conditions using sensor networks. In *Networking, Sensing and Control, 2004 IEEE International Conference on*, volume 2, pages 1008 – 1013 Vol.2, 2004.

[6] M. Crocco, A. Del Bue, and V. Murino. A bilinear approach to the position self-calibration of multiple sensors. *IEEE Transactions on Signal Processing*, 60(2):660–673, February 2012.

[7] Rita Cucchiara, Michele Fornaciari, Andrea Prati, and Paolo Santinelli. Mutual calibration of camera motes and rfids for people localization and identification. In *Proceedings of the Fourth ACM/IEEE International Conference on Distributed Smart Cameras*, ICDSC '10, pages 65–72, New York, NY, USA, 2010. ACM. ISBN 978-1-4503-0317-0.

[8] M. Ding, A. Terzis, I.J. Wang, and D. Lucarelli. Multi-modal calibration of surveillance sensor networks. In *IEEE Military Communications Conference (MILCOM 2006)*, pages 1–7. IEEE, 2006.

[9] M.G. Fugini, P. Maggiolini, C. Raibulet, and L. Ubezio. Risk management through real-time wearable services. In *Software Engineering Advances, 2009. ICSEA '09. Fourth International Conference on*, pages 163 –168, sept. 2009. doi: 10.1109/ICSEA.2009.33.

[10] A. Geiger, F. Moosmann, O. Car, and B. Schuster. Automatic camera and range sensor calibration using a single shot. In *Robotics and Automation (ICRA), 2012 IEEE International Conference on*, pages 3936 –3943, may 2012.

[11] T. Germa, F. Lerasle, N. Ouadah, and V. Cadenat. Vision and rfid data fusion for tracking people in crowds by a mobile robot. *Computer Vision and Image Understanding*, 114:641–651, June 2010. ISSN 1077-3142.

[12] Li Guan and Marc Pollefeys. A Unified Approach to Calibrate a Network of Camcorders and ToF cameras. In *Workshop on Multi-camera and Multi-modal Sensor Fusion Algorithms and Applications - M2SFA2 2008*, Marseille, France, 2008.

[13] K. Kanatani and D.D. Morris. Gauges and gauge transformations for uncertainty description of geometric structure with indeterminacy. *IEEE Transactions on Information Theory*, 47(5):2017–2028, 2001.

[14] K. Kanatani and Y. Sugaya. Factorization without factorization: complete recipe. *Mem. Fac. Eng. Okayama Univ*, 38(1&2):61–72, 2004.

[15] K. Kanatani, Y. Sugaya, and H. Ackermann. Uncalibrated factorization using a variable symmetric affine camera. *IEICE transactions on information and systems*, 90(5):851–858, 2007.

[16] A. Leykin and R. Hammoud. Robust multi-pedestrian tracking in thermal-visible surveillance videos. In *Computer Vision and Pattern Recognition Workshop, 2006. CVPRW'06. Conference on*, pages 136–136. IEEE, 2006.

[17] David Malan, Thaddeus Fulford-jones, Matt Welsh, and Steve Moulton. Codeblue: An ad hoc sensor network infrastructure for emergency medical care. In *In International Workshop on Wearable and Implantable Body Sensor Networks*, 2004.

[18] Gianluca Monaci, Pierre Vandergheynst, and Friedrich T. Sommer. Learning bimodal structure in audio-visual data. *IEEE Transactions on Neural Networks*, 20(12):1898–1910, 2009.

[19] P. Oguz-Ekim, J.P. Gomes, J. Xavier, and P. Oliveira. Robust localization of nodes and time-recursive tracking in sensor networks using noisy range measurements. *IEEE Transactions on Signal Processing*, (99):1–1, 2011.

[20] OpenNI. Openni framework@ONLINE, February 2012. URL http://www.openni.org/.

[21] C.J. Poelman and T. Kanade. A paraperspective factorization method for shape and motion recovery. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 19(3):206–218, 1997.

[22] M. Pollefeys and D. Nister. Direct computation of sound and microphone locations from time-difference-of-arrival data. In *Acoustics, Speech and Signal Processing, 2008. ICASSP 2008. IEEE International Conference on*, pages 2445 –2448, 2008.

[23] B. Shirmohammadi and C.J. Taylor. Distributed target tracking using self localizing smart camera networks. In *Proceedings of the Fourth ACM/IEEE International Conference on Distributed Smart Cameras*, pages 17–24. ACM, 2010.

[24] Gyula Simon, Miklós Maróti, Ákos Lédeczi, György Balogh, Branislav Kusy, András Nádas, Gábor Pap, János Sallai, and Ken Frampton. Sensor network-based countersniper system. In *Proceedings of the 2nd international conference on Embedded networked sensor systems*, SenSys '04, pages 1–12, New York, NY, USA, 2004. ACM. ISBN 1-58113-879-2. doi: 10.1145/1031495.1031497.

[25] P. Sturm and B. Triggs. A factorization based algorithm for multi-image projective structure and motion. *European Conference on Computer Vision (ECCV 1996)*, pages 709–720, 1996.

[26] Christopher Taylor, Ali Rahimi, Jonathan Bachrach, Howard Shrobe, and Anthony Grue. Simultaneous localization, calibration, and tracking in an ad hoc sensor network. In *Proceedings of the 5th international conference on Information processing in sensor networks*, IPSN '06, pages 27–33, New York, NY, USA, 2006. ACM. ISBN 1-59593-334-4. doi: 10.1145/1127777.1127785.

[27] C. Tomasi and T. Kanade. Shape and motion from image streams under orthography: a factorization method. *International Journal of Computer Vision*, 9(2):137–154, 1992.