Bi-objective Framework for Sensor Fusion in RGB-D Multi-View Systems: Applications in Calibration

Hassan Afzal, Djamila Aouada, Michel Antunes, David Fofi, Bruno Mirbach, Björn Ottersten

Bi-objective Framework for Sensor Fusion in RGB-D Multi-View Systems: Applications in Calibration

Hassan Afzal[1,*], Djamila Aouada[1], Michel Antunes[1], David Fofi[2], Bruno Mirbach[3], Björn Ottersten[1]

[1]Interdisciplinary Centre for Security, Reliability and Trust University of Luxembourg, 4, rue Alphonse Weicker, L-2721, Luxembourg
[2]Le2i - IUT Le Creusot, Université de Bourgogne,12 rue de la Fonderie, 71200, France
[3]IEE S.A., 11, rue Edmond Reuter, L-5326 Contern, Luxembourg
[*]Email: hassan.afzal@uni.lu

Abstract: Complete and textured 3D reconstruction of dynamic scenes has been facilitated by mapped RGB and depth information acquired by RGB-D cameras based multi-view systems. One of the most critical steps in such multi-view systems is to determine the relative poses of all cameras via a process known as extrinsic calibration. In this work, we propose a sensor fusion framework based on a weighted bi-objective optimization for refinement of extrinsic calibration tailored for RGB-D multi-view systems. The weighted bi-objective cost function, which makes use of 2D information from RGB images and 3D information from depth images, is analytically derived via the Maximum Likelihood (ML) method. The weighting factor appears as a function of noise in 2D and 3D measurements and takes into account the affect of residual errors on the optimization. We propose an iterative scheme to estimate noise variances in 2D and 3D measurements, for simultaneously computing the weighting factor together with the camera poses. An extensive quantitative and qualitative evaluation of the proposed approach shows improved calibration performance as compared to refinement schemes which use only 2D or 3D measurement information.

1. Introduction

RGB-D cameras provide simultaneous image and range data of the environment, offering enhanced sensing capabilities when compared to using single sensor modality. A non-exhaustive list of applications includes 3D telepresence systems [20], creation of viewpoint free 3D videos [19], simultaneous localization and mapping [17], or the acquisition of textured 3D surface models of static and dynamic scenes [26, 18, 34, 11].

The acquisition of complete and textured 3D models of scenes required in domains such as security and surveillance, health, and entertainment, can be accomplished by using two different approaches. The first consists of using a single moving RGB-D camera with its location constantly being tracked [17, 26, 34]. This solution is simple and attractive, however, has the drawback of not allowing to fully reconstruct dynamic scenes at each time-step. As an example, Dou et al. [13] presented a 3D scanning system for deformable objects using a single Kinect sensor. The scanning results are promising, remark that, however, the obtained 3D reconstructions are rigid, even if the objects in the environment were constantly moving and deforming. This issue can be solved by using multiple fixed RGB-D cameras covering the entire scene [18, 20, 11, 28]. In this case, the relative poses of all RGB-D cameras are required for aligning the partial 3D reconstructions. The

problem of estimating the relative poses of cameras in a multi-view system is known as extrinsic calibration.

Most of the works for extrinsic calibration of RGB-D multi-view systems rely on well established 2D camera based calibration routines [40, 10] and pose refinement procedures, e.g. Bundle Adjustment (BA) [36, 5, 11], using 2D feature points extracted from the RGB images [20, 8, 7]. The 3D information from the depth sensor has mainly been used in subsequent refinement steps using, e.g., the Iterative Closest Point (ICP) algorithm [32, 29, 38]. In this regard, the following question arises: how to optimally use both sources of complementary information.

In this paper, we investigate a strategy for RGB-D sensor fusion for the extrinsic calibration of multiple cameras. Instead of using 2D data from RGB images and 3D data from depth images independently, we propose a weighted bi-objective optimization scheme. We analytically derive a Least Squares (LS) based cost function, via the Maximum Likelihood (ML) method, that optimally combines the BA based 2D cost function with the ICP based 3D cost function. The sensor fusion is achieved by using a weighting factor that depends on two types of noise, one contaminating the 2D feature locations in the RGB images, and the second one contaminating the 3D point positions provided by the depth sensor. The experiments suggest that using the proposed joint cost for relative pose refinement provides more accurate results than the refinement schemes using 2D and 3D information separately.

In the absence of information regarding noise levels in the 2D and 3D feature points we propose an iterative scheme which simultaneously estimates the noise along with the estimation of calibration parameters. The proposed scheme is completely automated requiring no manual intervention and no heuristic parameter setting. The quantitative and qualitative experiments show that the proposed scheme is able to perform sensor fusion for accurate camera calibration without any prior information about noise characteristics.

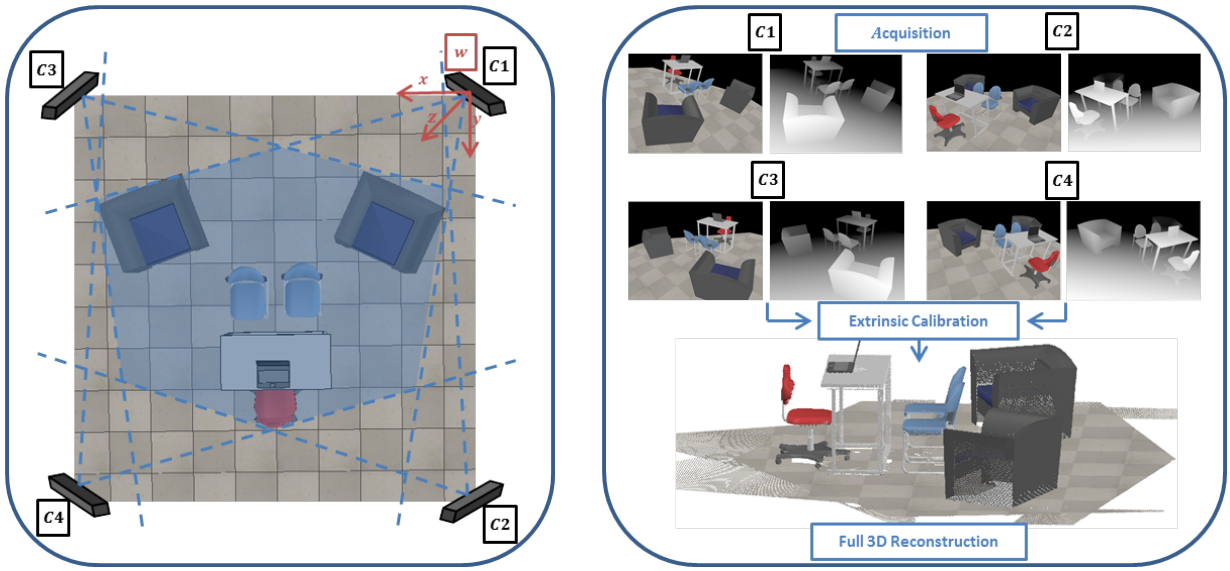
The present work extends and consolidates our previous work called *BAICP+* [3], which experimentally showed that in many cases, using a heuristically constructed weighted bi-objective refinement approach that combines 2D and 3D information provides better results than refinement approaches based on cost functions using only 2D or 3D information.

1.1. Contributions

- We present a sensor fusion framework based on weighted bi-objective optimization for refinement of extrinsic calibration of an RGB-D multi-view system. We derive an analytic expression for the weighting factor, in the bi-objective optimization, in terms of noise in measurements of RGB and depth sensors.
- We propose an iterative scheme for extrinsic calibration of an RGB-D multi-view system, which alternates between camera pose estimation, and the computation of measurement noise levels.
- We perform a thorough experimental evaluation on synthetic and real data, and show that fusing the RGB-D information using the proposed bi-objective optimization provides superior results when compared to refinement schemes that only use 2D or 3D feature information.

1.2. Article Overview

We start by giving a brief overview of state-of-the-art methods for extrinsic calibration used in RGB-D multi-view systems, as well as of bi-objective pose estimation approaches in Section 2. In



(a) RGB-D Multi-View System (4 cameras)

(b) Full 3D Scene Reconstruction

Fig. 1. *RGB-D Multi-View System with full scene 3D reconstruction in a simulated setup. (a) RGB-D Multi-View System (4 cameras) with field of view (FOV) of each camera. The highlighted region represents overlapping FOVs of all cameras. The global reference frame* `w` *is aligned with camera* ***C1****. (b) Steps required for Full 3D Scene Reconstruction using an RGB-D Multi-View System. Each camera acquires a RGB image and a depth image, which are used to estimate the relative pose of each camera with respect to* `w`*. After extrinsic calibration, estimated poses are used to put all acquisitions in* `w` *to get complete reconstruction.*

Section 3, the extrinsic calibration problem is formally presented. Section 4 gives a brief introduction of BA and ICP together with our previous work i.e., *BAICP+* [3], in which BA and ICP are heuristically combined. Section 5 analytically derives the expression for the weighted bi-objective cost function for refinement of extrinsic calibration parameters. Section 6 presents an automated iterative approach for camera pose estimation and estimation of the measurement noise parameters. In Section 7 and Section 8, we analyze and illustrate the benefits of the proposed approach via extensive experiments using synthetic and real data respectively. This is followed by a conclusion in Section 9.

2. Related Work

In this section we review state-of-the-art techniques for extrinsic calibration in an RGB-D multi-view system with a focus on the modality of data used. We also briefly overview the sensor fusion approaches, based on bi-objective optimization, for solving the pair-wise pose estimation problem.

A considerable amount of research has been, and is still being, carried out in the domain of RGB-D cameras based multi-view systems [6]. We are interested in analyzing the extrinsic calibration method used in such multi-view systems. Extrinsic calibration in a multi-view system requires information about the same points (feature points), in 3D space, to be acquired in different views. This information can be extracted from 2D or 3D camera acquisitions of objects with, e.g., known textural and/or geometrical properties [8, 18, 4, 22, 28]. A planar checkerboard pattern

first proposed by Zhang [40] is one of the most widely used objects for this purpose. Assuming correct mapping of RGB and depth sensors, Maimone and Fuchs [20], and Yang et al. [38] make use of such an object, with corners extracted from RGB and IR images respectively, to calibrate their systems with the help of Bougouet's calibration toolbox [10]. Berger et al. [7, 8], on the other hand, try to improve accuracy by calibrating all RGB and depth sensors together by extracting 2D feature points from both RGB and IR images with the help of a special checkerboard pattern consisting of diffuse and mirroring patches.

A major drawback of 2D only calibration approaches is their inability to tackle noise specific to depth sensors which causes problems in alignment of 3D data from multiple cameras. Various methods try to tackle this problem with the help of an explicit depth correction step based on comparing known and measured depths for each camera, e.g., [20, 18, 25, 11] etc. There are other methods, such as [38], which add a final refinement step based on the ICP algorithm, using 3D data only, which tries to mitigate the pose misalignment due to depth specific sensor noise. Penelle et al. [29] propose to use only 3D points, corresponding to 2D feature points (from checkerboard) extracted from RGB images, in the ICP based calibration scheme where initial estimates are provided via the RANSAC algorithm. Nakazawa et al. [25], on the other hand, use 2D feature points in the BA algorithm for pose refinement but perform corrections using alignment of 3D data as well. Miller et al. [22] propose to use 3D information related to foreground objects to obtain an occlusion aware energy minimization scheme for auto-calibration. Deng et al. [35], use 3D feature points observed at different locations in the scene to compute a smooth field of rigid transformation instead of a single rigid transformation to align the measurements of two RGB-D cameras.

Dou and Fuchs [11], in their work on multi-view 3D reconstruction, proposed to combine 2D and 3D information in a weighted bi-objective optimization scheme derived from their previous work on pair-wise pose tracking for mono-view 3D reconstruction [12]. They propose to use matching feature points extracted via Scale-Invariant Feature Transform (SIFT) from RGB images with matching planes extracted from 3D/depth images in a weighted bi-objective BA scheme. The weighting factor is selected empirically for all experiments. A similar approach is proposed by Henry et al. [17], using a global ICP scheme to align 2D visual feature points and 3D/depth measurements from multiple views but the weights are, again, selected empirically. Tykkala et al. [37] use what they call an image based direct ICP approach for pairwise pose estimation. They propose to compute the weighting factor via a heuristic measure using ratio of the median intensity and the depth values of selected points. Michot et al. [21] propose to use a weighted bi-objective BA scheme for the multi-sensor Simultaneous Localization and Matching (SLAM) problem. They discuss the dependence of the weighting factor on the ratio of the noise variance for each sensor's measurement and formulate their bi-objective optimization by using a Mean Squared Error (MSE) based cost function from individual sensors. They investigate three methods for automatic weight computation namely L-Curve, L-Tangent Norm and cross validation with experiments showing that the L-Curve based method performs better than the others.

This brief survey shows that most work done on extrinsic calibration of RGB-D multi-view systems is based on calibration schemes which use 2D or 3D information independently. These methods do not take into account the relative accuracy of 2D and 3D measurement in a systematic manner to, e.g., give more importance to less noisy measurements.

In this work, we perform sensor fusion by formulating a weighted bi-objective optimization scheme based on 2D and 3D cost functions for performing global refinement of poses in RGB-D multi-view systems. We draw upon work in the Robotics domain [21, 37, 15, 12], where several sensor modalities are exploited for pair-wise pose estimation. The proposed bi-objective optimiza-

tion uses cost functions from both the BA and ICP algorithms, based on 2D and 3D measurements, in a single unified cost function. The key to combining the two cost functions is the information about noise in the 2D and 3D measurements which is reflected in the weighting factor. In the absence of this information we propose to use a simple approach which iteratively estimates the noise parameters given the current camera poses and vice versa. We show the validity of the proposed approach by achieving improved results in the experiments performed under different conditions.

3. Extrinsic Calibration Problem of Multiple RGB-D Sensors

Notation: The following notation will be adopted. Subscripts indicate camera or reference frame indexes and superscripts indicate point indexes:

$\mathtt{w}$:	world reference frame.
$\mathbf{A}$:	matrix.
$\mathbf{p}$:	vector.
l, M :	scalars.
$tr(\mathbf{A})$:	trace of matrix $\mathbf{A}$.
$\mathbf{A}^\intercal$:	transpose of matrix $\mathbf{A}$.
$\psi(\cdot)$, :	$\mathtt{w}$ to image plane projection.
$\hat{\mathbf{p}}$, $\hat{\mathbf{T}}_l$:	estimates of $\mathbf{p}$ and $\mathbf{T}_l$, respectively.
$\mathbf{I}_n$:	identity matrix of dimension $n \times n$.
$\mathbf{0}_n$:	null vector of dimension $n \times 1$.

In this section, we will formulate the extrinsic calibration problem for an RGB-D multi-view system. Let us consider a multi-view system composed of N, intrinsically calibrated, RGB-D cameras with intersecting FOVs, as shown in Fig. 1. Every RGB-D camera l, with $l = 1, \cdots, N$, acquires an RGB image $\mathbf{C}_l$ and a 3D vertex map $\mathbf{V}_l$, with associated known matrix of intrinsic parameters $\mathbf{K}_l$.

In order to correctly align the partial 3D reconstructions $\{\mathbf{V}_l\}$, where $l = 1, \cdots, N$, acquired by N RGB-D cameras, it is necessary to accurately estimate their positions with respect to a global reference frame, referred to as `world` and denoted by $\mathtt{w}$, as shown in Fig. 1. Each camera's relative position with respect to $\mathtt{w}$ is defined by:

$$\mathbf{T}_l = \begin{pmatrix} \mathbf{R}_l & \mathbf{t}_l \\ \mathbf{0}_3^\intercal & 1 \end{pmatrix}, \tag{1}$$

where $\mathbf{T}_l \in SE(3)$ represents the rigid transformation, from camera l to $\mathtt{w}$. The matrix $\mathbf{R}_l$ is rotation matrix in $SO(3)$ and $\mathbf{t}_l \in \mathbb{R}^3$ is translation vector. Therefore the same point $\mathbf{p} \in \mathbb{R}^3$ in $\mathtt{w}$ viewed by camera l as $\mathbf{p}_l$ and by cameras k as $\mathbf{p}_k$ can be related to the cameras' reference frames as follows:

$$\mathbf{R}_l \mathbf{p}_l + \mathbf{t}_l = \mathbf{R}_k \mathbf{p}_k + \mathbf{t}_k. \tag{2}$$

Similarly, for a given point $\mathbf{x} \in \mathbb{R}^3$ in $\mathtt{w}$, its projection on each camera's image plane results in 2D pixel coordinates $\mathbf{q}_l$, such that:

$$\mathbf{q}_l = \psi\left(\mathbf{K}_l, \mathbf{T}_l, \mathbf{x}\right), \quad \forall l, \tag{3}$$

where $\psi(.)$ is `world` to image plane projection function.

Let us assume that all cameras are of resolution M. The color pixel positions in $\mathbf{C}_l$ may be represented by the points $\mathbf{q}_l^i \in [\mathbf{q}_l^1, \cdots, \mathbf{q}_l^M]$ where $i \in \{1, \cdots, M\}$. Similarly, the 3D coordinates in $\mathbf{V}_l$ may be represented by points $\mathbf{p}_l^k \in [\mathbf{p}_l^1, \cdots, \mathbf{p}_l^M]$ where $k \in \{1, \cdots, M\}$.

The problem at hand may therefore be stated as follows. Given N RGB-D cameras in a multi-view system with acquired RGB images $\{\mathbf{C}_1, \cdots, \mathbf{C}_N\}$ and 3D vertex maps $\{\mathbf{V}_1, \cdots, \mathbf{V}_N\}$, we assume knowledge of $H \leq M$ matching points in each camera's RGB image plane referred to as 2D feature points and denoted as $[\mathbf{q}_l^1, \cdots, \mathbf{q}_l^H]$. Similarly, we assume knowledge of $J \leq M$ matching 3D points in each camera's 3D vertex map called 3D feature points and denoted as $[\mathbf{p}_l^1, \cdots, \mathbf{p}_l^J]$. Moreover we assume knowledge of each camera's intrinsic parameters, $\mathbf{K} = [\mathbf{K}_1, \cdots, \mathbf{K}_N]$. Using this information, we want to find the estimates of the parameters $\mathbf{T} = [\mathbf{T}_1, \cdots, \mathbf{T}_N]$.

4. Background

In this section, we introduce the two pose refinement algorithms namely BA and ICP, from which we derive the 2D and 3D cost functions used in our proposed bi-objective cost function. We also introduce *BAICP+* [3], in which we proposed a bi-objective pose refinement scheme, but, with heuristically defined weight and scaling parameters.

4.1. Bundle Adjustment (BA)

For refinement of extrinsic calibration parameters using 2D feature points only, we use a cost function from the Bundle Adjustment (BA) algorithm [36]. It has been the method of choice for problems related to multi-view 3D reconstruction and pose refinement based on 2D feature points extracted from RGB images [36]. Bundle Adjustment (BA) requires an initial estimate of the pose parameters. Moreover, it also requires an estimate of 3D points i.e., $[\mathbf{x}^1, \cdots, \mathbf{x}^H]$, corresponding to available 2D feature points $[\mathbf{q}_l^1, \cdots, \mathbf{q}_l^H]$. These estimates are then refined by computing the error of projection of estimate of each 3D point $\mathbf{x}^h$, $h = 1, \cdots, H$, corresponding to the 2D feature point $\mathbf{q}_l^h$ to camera l via:

$$\mathbf{a}_l^h(\mathbf{S}_l^h) = \mathbf{q}_l^h - \psi\left(\mathbf{K}_l, \mathbf{T}_l, \mathbf{x}^h\right), \tag{4}$$

where $\mathbf{a}_l^h(\mathbf{S}_l^h) \in \mathbb{R}^2$ and $\mathbf{S}_l^h = \left(\mathbf{T}_l, \mathbf{x}^h\right)$ [1]. Therefore, the total BA cost to be minimized for the refinement of estimates of each camera's pose parameters together with the estimates of 3D points corresponding to 2D feature points is given as:

$$V_{BA}(\mathbf{S}) = \sum_{l=1}^{N} tr(\mathbf{A}_l^\intercal(\mathbf{S}_l)\mathbf{A}_l(\mathbf{S}_l)), \tag{5}$$

where $\mathbf{S} = (\mathbf{T}, \mathbf{X})$, $\mathbf{S}_l = (\mathbf{T}_l, \mathbf{X})$, $\mathbf{X} = [\mathbf{x}^1, \cdots, \mathbf{x}^H]$ and $\mathbf{A}_l(\mathbf{S}_l) = [\mathbf{a}_l^1(\mathbf{S}_l^1), \cdots, \mathbf{a}_l^H(\mathbf{S}_l^H)]$.

4.2. Iterative Closest Point (ICP)

For refinement of the extrinsic calibration parameters using 3D feature points only, we use the cost function from the global Iterative Closest Points (ICP) algorithm [9]. ICP algorithm has been the de facto solution for pose refinement problems when only 3D feature points are available [27, 9]. ICP algorithm also uses initial estimates of the pose parameters and minimizes the Euclidean distance

[1] BA can also refine the estimate of intrinsics $\mathbf{K}_l$ if required

between corresponding 3D feature points from different views, such that:

$$\mathbf{b}^j_{l,k}(\mathbf{T}_l, \mathbf{T}_k) = (\mathbf{R}_l \mathbf{p}^j_l + \mathbf{t}_l) - (\mathbf{R}_k \mathbf{p}^j_k + \mathbf{t}_k), \tag{6}$$

where $\mathbf{b}^j_{l,k}(\mathbf{T}_l, \mathbf{T}_k) \in \mathbb{R}^3$ and $j \in [1, \cdots, J]$. Therefore, the total ICP cost to be minimized for refinement of each camera's pose parameters is given as:

$$V_{ICP}(\mathbf{T}) = \sum_{\substack{1<l,k<N \\ l \neq k}} tr(\mathbf{B}^\intercal_{l,k}(\mathbf{T}_l, \mathbf{T}_k)\mathbf{B}_{l,k}(\mathbf{T}_l, \mathbf{T}_k)), \tag{7}$$

where $\mathbf{B}_{l,k}(\mathbf{T}_l, \mathbf{T}_k) = [\mathbf{b}^1_{l,k}(\mathbf{T}_l, \mathbf{T}_k), \cdots, \mathbf{b}^J_{l,k}(\mathbf{T}_l, \mathbf{T}_k)]$.

4.3. BAICP+

In our previous work [3], we used the information provided by RGB-D cameras in a bi-objective optimization scheme for extrinsic calibration refinement. It combines the BA and ICP cost functions based on 2D and 3D feature points respectively resulting in the cost function for *BAICP+*:

$$V_{BAICP}(\mathbf{S}) = \frac{(1-c)}{a} V_{ICP}(\mathbf{T}) + \frac{sc}{b} V_{BA}(\mathbf{S}), \tag{8}$$

where $c \in [0, 1]$ is the weighting factor and $s = (\frac{avgDepth}{avgFocal})^2$ is a heuristic scaling factor, based on the ratio of the average depth of points in $\mathbf{x}$ versus average focal length across all views. The parameters a and b denote the total number of 3D point correspondences and 2D feature points points across all views.

5. Bi-Objective Extrinsic Calibration

In this section, we present the bi-objective optimization for refinement of the extrinsic calibration parameters in an RGB-D multi-view system. We use cost functions defined in the previous section which use 2D and 3D feature points extracted from RGB images and vertex maps, respectively.

In this work, we propose to formally analyze and derive an expression for the cost function, based on ML estimations, of the bi-objective optimization taking into account the noise affecting both 2D and 3D measurement/feature points. We assume the presence of independent additive Gaussian noise in each coordinate of the 3D feature points such that:

$$\tilde{\mathbf{p}}^j_l \sim \mathcal{N}\left(\mathbf{p}^j_l, \sigma^2_{3D}\mathbf{I}_3\right), \tag{9}$$

where $\tilde{\mathbf{p}}^j_l$ is the noisy 3D point and $\mathbf{p}^j_l$ is the noise free point. Similarly for 2D feature points we have:

$$\tilde{\mathbf{q}}^h_l \sim \mathcal{N}(\mathbf{q}^h_l, \sigma^2_{2D}\mathbf{I}_2), \tag{10}$$

where $\tilde{\mathbf{q}}^h_l$ is the noisy 2D point and $\mathbf{q}^h_l$ is the noise free point. This means that we have to use the noisy 2D and 3D feature points to estimate the pose parameters. This leads to redefining the 3D error function $\mathbf{b}^j_{l,m}(\mathbf{T}_l, \mathbf{T}_m)$, given in (6), such that it computes the error between noisy points $\tilde{\mathbf{p}}^j_l$ and $\tilde{\mathbf{p}}^m_l$ projected to $\mathtt{w}$, from camera l and m, using the pose parameters $\mathbf{T}_l$ and $\mathbf{T}_m$, respectively. Similarly the 2D error function $\mathbf{a}^h_l(\mathbf{S}^h_l)$, given in (4) where $\mathbf{S}^h_l = (\mathbf{T}^l, \mathbf{x}^h)$, is redefined such that it

computes the 2D error between back projection of the estimated 3D point $\mathbf{x}^h$ to camera l, using $\mathbf{T}_l$ and $\mathbf{K}_l$, and the corresponding noisy 2D feature point $\tilde{\mathbf{q}}_l^h$.

Now, we can define the distribution the 3D error $\mathbf{b}_{l,m}^j(\mathbf{T}_l, \mathbf{T}_m)$ is drawn from by considering the noise free 3D points $\mathbf{p}_l^j$ and $\mathbf{p}_m^j$ in (2) such that [32]:

$$\begin{aligned}\mathbf{b}_{l,m}^j(\mathtt{T}_l, \mathtt{T}_m) &\sim \mathcal{N}\left((\mathbf{R}_l\mathbf{p}_l^j + \mathbf{t}_l) - (\mathbf{R}_m\mathbf{p}_m^j + \mathbf{t}_m), \mathbf{R}_l\sigma_{3D}^2\mathbf{I}_3\mathbf{R}_l^\intercal + \mathbf{R}_m\sigma_{3D}^2\mathbf{I}_3\mathbf{R}_m^\intercal\right) \\ &= \mathcal{N}(\mathbf{0}_3, 2\sigma_{3D}^2\mathbf{I}_3). \end{aligned} \tag{11}$$

Similarly, considering the noise free 2D measurements in (3), we have $\mathbf{a}_l^h(\mathbf{S}_l^h) \sim \mathcal{N}(\mathbf{0}_2, \sigma_{2D}^2\mathbf{I}_2)$. It is clear from (11) that since $\mathbf{b}_{l,m}^j(\mathbf{T}_l, \mathbf{T}_m)$, which is based on the ICP algorithm, uses two noisy 3D feature points, hence, the variance of the corresponding distribution is two times the variance of noise in each 3D feature point. This is in contrast to the variance of distribution corresponding to $\mathbf{a}_l^h(\mathbf{S}_l^h)$, which is based on the BA algorithm and uses only one noisy 2D feature point [21].

Using $\mathbf{b}_{l,m}^j(\mathbf{T}_l, \mathbf{T}_m)$ and $\mathbf{a}_l^h(\mathbf{S}_l^h)$, we want to find the likelihood cost function, maximum of which gives the Maximum Likelihood Estimate (MLE) of the parameters $\mathbf{S} = (\mathbf{T}, \mathbf{x})$. Since the MLE with Gaussian model is equivalent to the Least Squares Estimate (LSE) [30], we can directly get:

$$\hat{\mathbf{S}} = \arg\min_{\mathbf{S}} \sum_{\substack{1<l,m<N \\ l\neq m}} \frac{1}{2\sigma_{3D}^2} tr\left(\mathbf{B}_{l,m}^\intercal(\mathbf{T}_l, \mathbf{T}_m)\mathbf{B}_{l,m}(\mathbf{T}_l, \mathbf{T}_m)\right) + \sum_{l=1}^{N} \frac{1}{\sigma_{2D}^2} tr\left(\mathbf{A}_l^\intercal(\mathbf{S}_l)\mathbf{A}_l(\mathbf{S}_l)\right). \tag{12}$$

Therefore, the total cost to be minimized is:

$$V(\mathbf{S}) = V_{ICP}(\mathbf{T}) + wV_{BA}(\mathbf{S}), \tag{13}$$

where $w = \frac{2\sigma_{3D}^2}{\sigma_{2D}^2}$ is the weighting factor. The cost function in (12) optimally combines information from RGB and depth sensors, to be used in the pose refinement scheme, by taking into account the noise levels in the 2D and 3D points. It formally defines the the relationship of measurement noise in the 2D and 3D feature points with the weighting factor w. In case the assumption of noise with same variances affecting all 2D and 3D points respectively, does not hold and information about the noise variances affecting each point is available, it can be incorporated in the proposed framework. Moreover, the use of the ICP based cost also allows the use of all the 3D points acquired by each sensor (with the help of nearest neighbor correspondence) in the optimization scheme when only 2D feature points are available.

The cost function (13) is a non-linear function of the parameters $\mathbf{S}$ and we resort to numerical search methods [24] to optimize the criterion. Please refer to Appendix 11 for further discussion.

6. Weighting Factor Estimation

In this section we discuss the automatic and simultaneous estimation of the weighting factor w in (13), together with the camera poses in the absence of information regarding noise affecting both the 2D and 3D measurements. We propose an approach which alternates between camera pose estimation and estimation of the 2D and the 3D noise variances to arrive at a suitable solution.

In the previous section, the estimates of camera pose parameters and 3D points in $\mathtt{w}$ corresponding to 2D feature points, were computed based on known 2D and 3D feature points and the noise

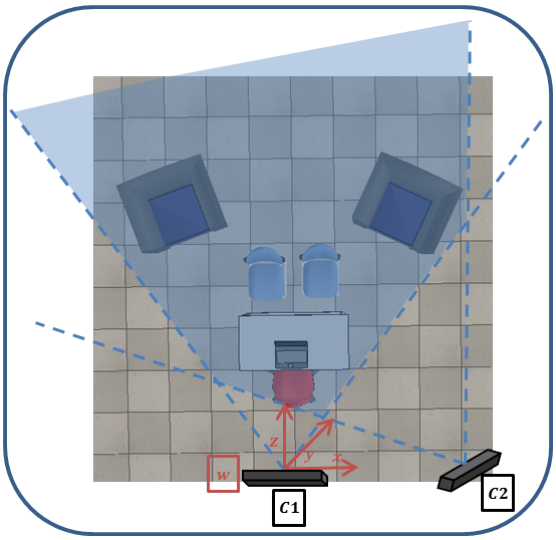


Fig. 2. *RGB-D Multi-View System (2 cameras) with field of view (FOV) of each camera. The highlighted region represents overlapping FOVs of all cameras. The global reference frame* `w` *is aligned with camera* ***C1****.*

affecting them. We assumed the presence of Gaussian noise with zero mean and variances of σ_{2D}^2 and σ_{3D}^2 in 2D and 3D measurements, respectively. These parameters, in turn, define the weighting factor w which is instrumental in constructing the sensor fusion framework by optimally combining the 2D and 3D cost functions to estimate the camera poses. In real-world scenarios, however, information about the noise affecting one or both sensor measurements is often unavailable. This makes the computation of a correct w difficult. As mentioned in Section 2, researchers have tried to estimate the optimal weighting factor, for their proposed bi-objective schemes, for solving mainly the pair-wise pose estimation problem. The commonly used used methods range from using simple heuristic measures such as in the case of [37] to more complex methods, based on analysis of trade-off between residuals of two cost functions and based on learning via cross-validation, such as in the case of [21].

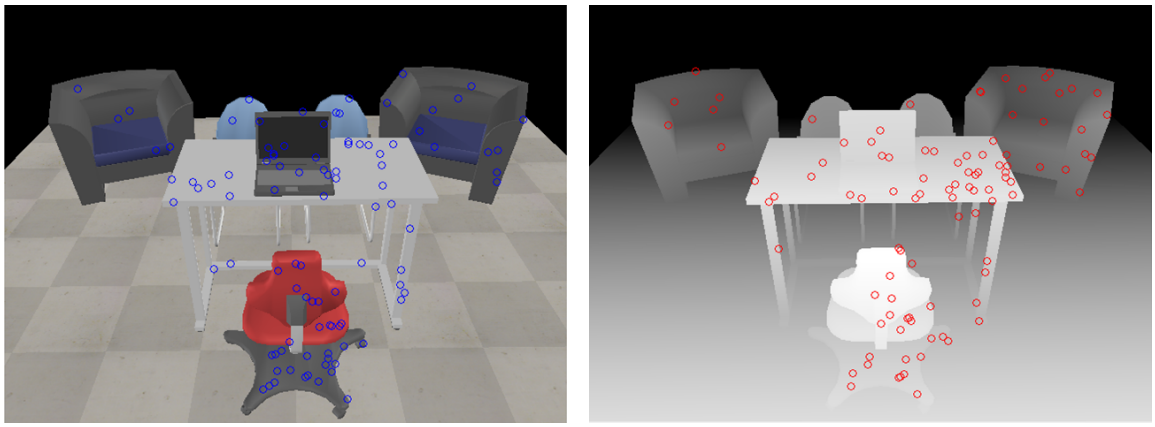
In this work, we propose to use a simple method for automatic estimation of the weighting factor w which finds its basis in finding the MLE of noise variances, σ_{2D}^2 and σ_{3D}^2, using the 2D and 3D feature points together with the current estimates of camera poses and 3D points in $\hat{\mathbf{S}}$. The MLE of the variance σ_{3D}^2 is given as [30]:

$$\hat{\sigma}_{3D}^2 = \sum_{\substack{1<l,m<N \\ l \neq m}} \frac{tr(\mathbf{B}_{l,m}^{\intercal}(\mathbf{T}_l, \mathbf{T}_m)\mathbf{B}_{l,m}(\mathbf{T}_l, \mathbf{T}_m))}{2a}, \tag{14}$$

where a is the total number of 3D feature correspondences across all views. Similarly, the MLE of the variance σ_{2D}^2 is computed via:

$$\hat{\sigma}_{2D}^2 = \sum_{l=1}^{N} \frac{tr(\mathbf{A}_l^{\intercal}(\mathbf{S}_l)\mathbf{A}_l(\mathbf{S}_l))}{b}, \tag{15}$$

where b is the total number of 2D feature points found across all views.



(a) Features extracted from RGB image

(b) Features extracted from depth image

Fig. 3. *Features extracted from RGB and depth images of camera **C1** in the multi-view system composed of 2 cameras as shown in Fig. 2. The extracted feature points are also visible to camera **C2**.*

We follow an iterative approach whereby using 2D and 3D feature points and an initial estimate $\hat{\mathbf{S}}$, the MLE estimates of noise variances and hence of w are obtained via (14) and (15). This initial estimate of w is then used to find an updated estimate of $\mathbf{S}$ using (13) via non-linear optimization which, in turn, is used to update the estimate of w. This process is repeated for a fixed number of iterations until the estimates of $\mathbf{S}$ and w converge.

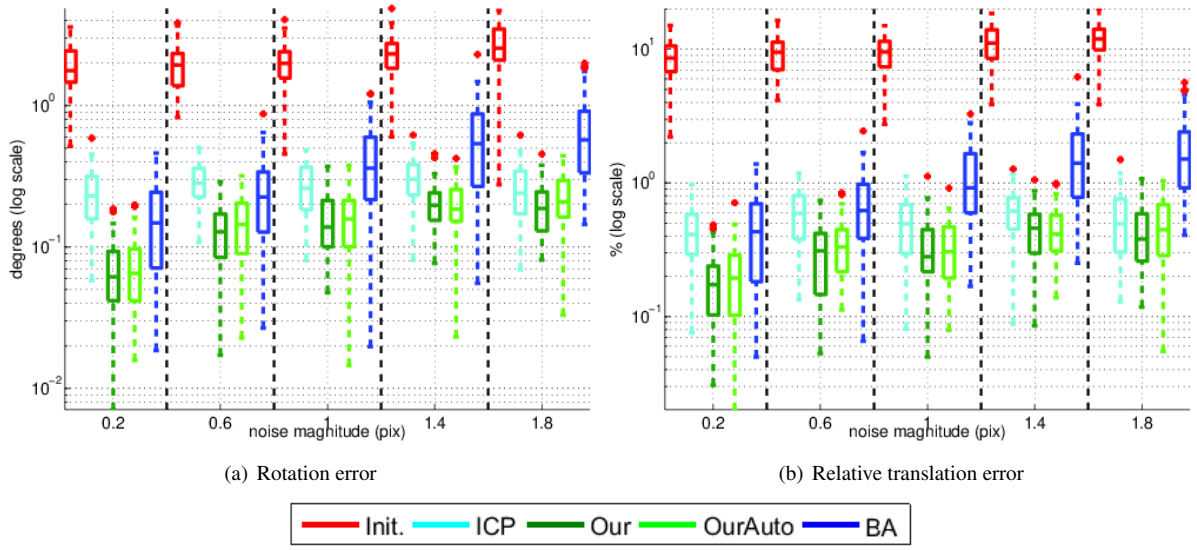
7. Experiments with Synthetic Data

In this section, we carry out a quantitative performance analysis of the proposed bi-objective refinement with a known and an unknown weighting factor.

7.1. Evaluation Methodology and Parameters

We use V-REP [1] to simulate 2 and 4 cameras based RGB-D multi-view systems, with overlapping FOVs, as shown in Fig. 2 and Fig. 1, respectively. In both cases, the global reference frame $\mathtt{w}$ lies in camera ***C*1**. We simulate a scene containing several objects such as chairs, a table, sofas etc. The acquired noise-free data, in the form of RGB and depth images, is assumed to be perfectly mapped in each camera's RGB sensor's reference frame with known intrinsics. After data acquisition, random points, visible to all cameras, are extracted as feature points in both RGB and depth images as shown in Fig. 3 (points on the floor are discarded). Features extracted from depth maps are converted to the corresponding 3D points via known intrinsics.

In the next step, noise is added to the extracted 2D and 3D feature points. We assume the presence of independent Gaussian noise in each coordinate of position of 2D feature points with zero mean and standard deviation σ_{2D} similar to [39]. The value of σ_{2D} is varied between 0.2 to 1.8 pixels with a step size of 0.4 pixels. Depth sensor measurements in RGB-D cameras suffer from different types of systematic and non-systematic errors as investigated in [33, 14]. For our scheme we propose to counter, beforehand, the systematic errors in depth measurements of each camera via a correction step, based on comparing known and measured depths [20, 25]. Therefore, for all remaining errors we assume the presence of additive independent Gaussian noise in each

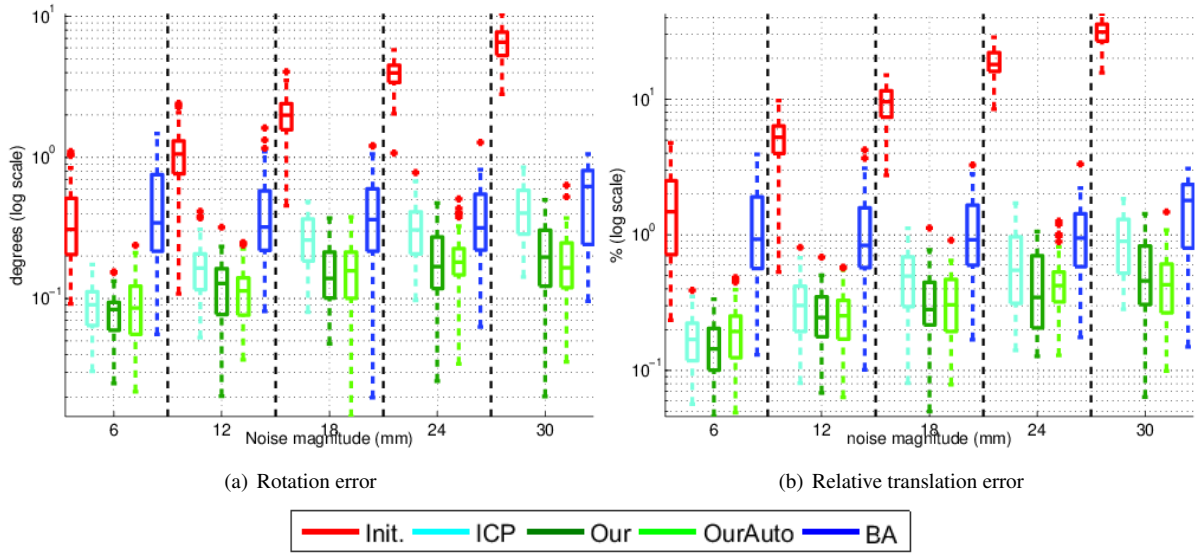


(a) Rotation error (b) Relative translation error

Fig. 4. *Error distribution of pose estimates for camera* ***C2*** *in a two camera setup. 100 2D and 100 3D feature points are used. The following methods are compared: Init. - Initial pose obtained using a DLT like approach (2D feature points and corresponding 3D points are used) [10, 16]; ICP - refinement of Init. using Iterative Closest Point (only 3D feature points are used); Our - refinement of Init. using our bi-objective optimization with known w (2D and 3D feature points are used); OurAuto - refinement of Init. using our bi-objective optimization with unknown w (2D and 3D feature points are used); BA - refinement of Init. using Bundle adjustment (only 2D feature points are used). Gaussian noise is added to the data, being the variance of the 3D noise fixed ($\sigma_{3D} = 18$mm), and the 2D noise σ_{2D} is varied between 0.2 and 1.8 pixels (horizontal axes).*

coordinate of 3D feature points in each view with zero mean and standard deviation σ_{3D}. The value of σ_{3D} is varied between 6 to 30 mm with a step size of 6 mm to keep it in the range of errors computed in [14]. We test the performance of the proposed scheme under various conditions by varying the number of cameras and their positions as shown in Fig. 2 and Fig. **??**, by varying the noise magnitude in 2D and 3D feature points as explained above, and by varying the number of 2D and 3D feature points. For each configuration, 50 noise realizations are generated. For each noise realization, 2D feature points and their corresponding noisy 3D measurements from vertex maps are used to initialize the pose estimates via a Direct Linear Transform (DLT) based approach [10, 16]. Using the initial pose estimates, optimization is carried out via the proposed scheme, with known noise parameters as explained in Section 5, and with unknown noise parameters using the automatic iterative estimation scheme as explained in Section 6 (required 3 iterations to converge in most cases). Furthermore, optimization is also carried out via ICP algorithm using 3D feature points only, and via BA algorithm using 2D feature points only.

Accuracy of the estimated poses is computed by comparison with the ground truth poses as done in [39]. Two measures of accuracy are computed. First is the angular magnitude of residual rotation computed via $\hat{\mathbf{R}}_l^T \mathbf{R}_l$, and second is the relative translation error which is computed via $\frac{\|\hat{\mathbf{t}}_l - \mathbf{t}_l\|}{\|\mathbf{t}_l\|}$. The results of 50 realizations showing the accuracy, of each initialization and of each refinement approach, for each configuration are plotted by using the function *boxplot* in MATLAB as shown in Fig. 4 - 12. The horizontal line inside each box marks the median, the edges mark the 25th and the 75th percentiles, the whisker edges show most extreme data points with outliers



(a) Rotation error (b) Relative translation error

Fig. 5. *Error distribution of pose estimates for camera* ***C2*** *in a two camera setup. 100 2D and 100 3D feature points are used. Gaussian noise is added to the data, being the variance of the 2D noise fixed ($\sigma_{2D} = 1pix$), and the 3D noise σ_{3D} is varied between $8mm$ and $30mm$ (horizontal axes).*

plotted separately as red crosses.

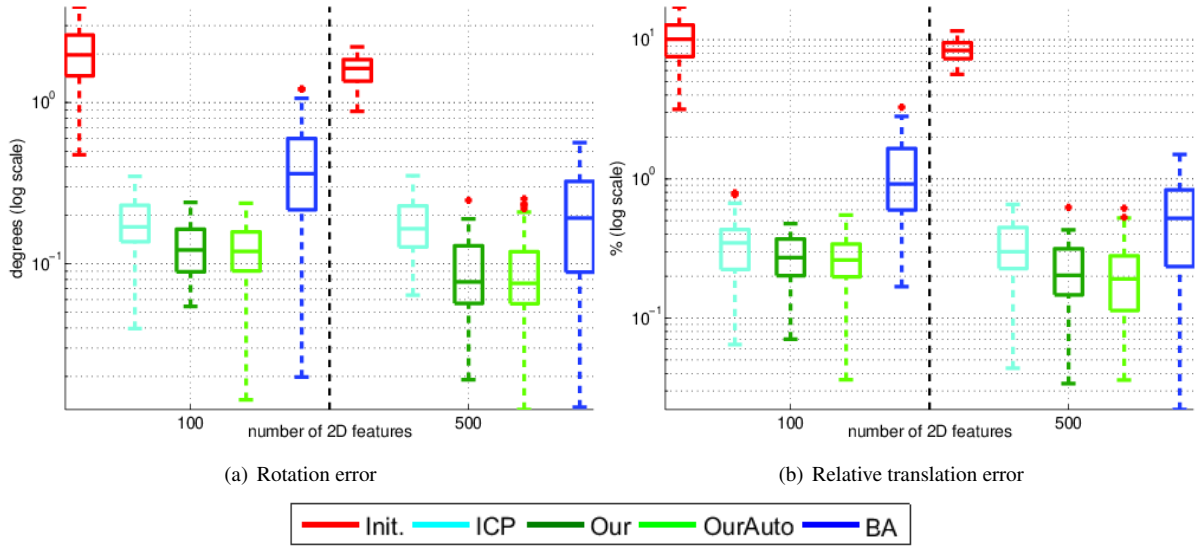
The implementation of the proposed bi-objective optimization scheme and ICP is based on the non-linear optimization via Levenberg Marquardt (LM) algorithm [23], while the implementation of BA is based on a sparse variant of the LM algorithm called Sparse Bundle Adjustment (SBA) [31, 36].

7.2. System Composed of Two Sensors

This section compares the performance of the proposed bi-objective optimization scheme, with known and unknown weighting factor, ICP and BA for refinement of camera pose parameters in a two camera setup shown in Fig. 2. The pose of camera ***C2*** with respect to camera ***C1*** is estimated. After initialization, pose refinement is carried out using the four refinement methods and results are plotted in Fig. 4 - 8.

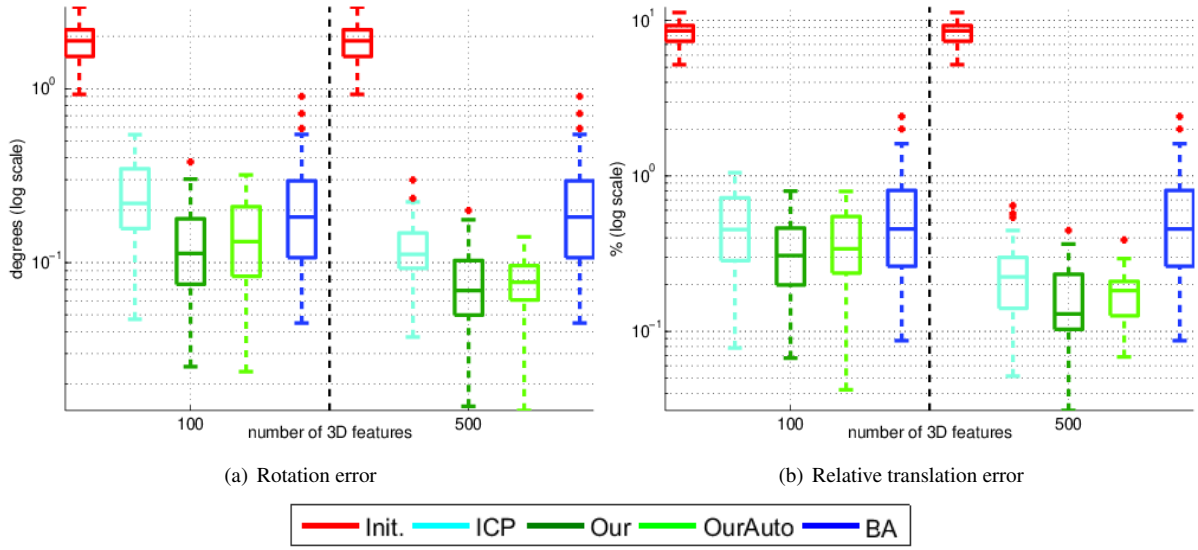
7.2.1. Varying Noise Levels: In this experiment, the extrinsic calibration is carried out using 100 2D feature points and 100 3D feature points. Fig. 4 shows the error distribution for fixed 3D noise and varying 2D noise, while Fig. 5 shows the distribution in case the 2D noise is kept fixed, and the 3D noise is varied.

As expected, the accuracy of the extrinsic calibration decreases with increasing noise levels. Also, all pose refinement approaches are able to improve the initial pose estimates, explained by the fact that only inlier data points are generated (no wrong matching feature points are included). A careful analysis of the results shows that our bi-objective optimization scheme with known w, which uses simultaneously the 2D and 3D data, provides better pose estimations when compared to ICP and BA, where only 3D feature points and 2D feature points are used, respectively. The weighting factor based on the noise variance information in (13) automatically gives prominence to more reliable data, decreasing the impact of the other sensor modality. Moreover, it shows that our proposed automatic iterative estimation scheme used in the absence of information regarding



(a) Rotation error (b) Relative translation error

Fig. 6. *Error distribution of pose estimates for camera* ***C2*** *in a two camera setup. Gaussian noise is added to the data ($\sigma_{2D} = 1pix$, $\sigma_{3D} = 18mm$), 250 3D feature points and a varying number of 2D feature points (horizontal axes) is considered.*

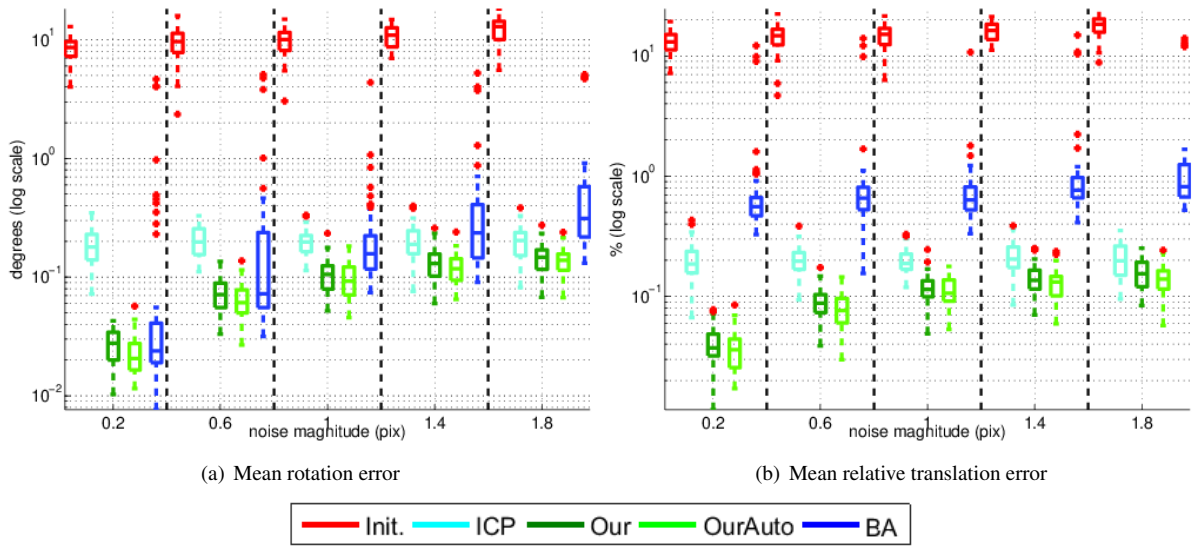


(a) Rotation error (b) Relative translation error

Fig. 7. *Error distribution of pose estimates for camera* ***C2*** *in a two camera setup. Gaussian noise is added to the data ($\sigma_{2D} = 1pix$),$\sigma_{3D} = 18mm$), 250 2D feature points and a varying number of 3D feature points (horizontal axes) is considered.*

noise parameters, and hence unknown w, is robust and also more accurate when compared to BA and ICP, and in most cases nearly as accurate as the method with known w.

7.2.2. Varying Data Ratio: In this experiment, the extrinsic calibration is carried out using fixed noise variance ($\sigma_{2D} = 1$pix, $\sigma_{3D} = 18$mm). Fig. 6 shows the error distribution for a fixed number of 3D points and a varying number of 2D points, while Fig. 7 shows the distribution in



(a) Mean rotation error (b) Mean relative translation error

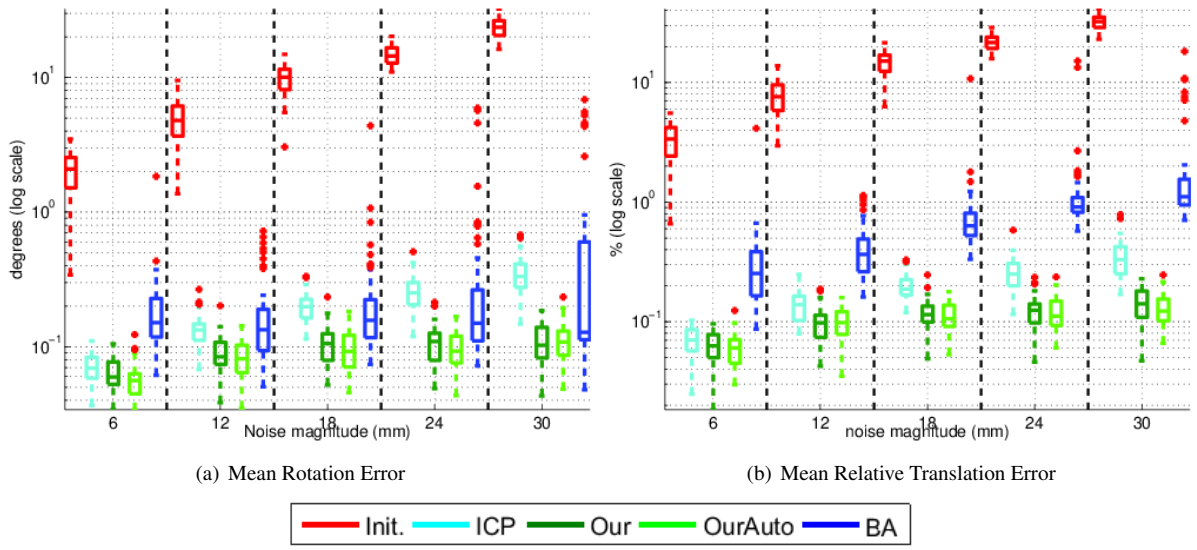
Fig. 8. *Mean error distribution, of pose estimates for cameras **C2**, **C3** and **C4**. in a four camera setup. 100 2D and 100 3D feature points are used. Gaussian noise is added to the data, being the variance of the 3D noise fixed ($\sigma_{3D} = 18mm$), and the 2D noise σ_{2D} is varied between 0.2 and 1.8 pixels (horizontal axes).*

case the 2D points are kept fixed, and the number of 3D points is varied. Since the initial poses are obtained by using 2D feature points and their corresponding 3D points, the initialization varies in Fig. 6 as number of 2D feature points vary but stays approximately the same in Fig. 7 as the number of 2D feature points remain fixed. The conclusions drawn in the previous section regarding improved accuracy of the proposed approaches hold, and these results show that the proposed scheme generalizes for different ratios between the number of 2D and 3D points. Increasing the number of data points of one of the sensor modalities always improves the extrinsic calibration accuracy for the algorithms using those modalities. Moreover, the results in Fig. 4, Fig. 5, Fig. 6, and Fig. 7 show the increased robustness of the proposed approach and ICP to bad initialization as compared to BA.

7.3. System Composed of Four Sensors

This section compares performance of the proposed bi-objective optimization scheme with ICP and BA for refinement of camera pose parameters in a four camera setup shown in Fig. **??**. The poses of cameras ***C2***, ***C3*** and ***C4*** are aligned with camera ***C1***. After initialization, pose refinement is carried out using the four refinement methods and results are plotted.

7.3.1. Varying Noise Levels: In this experiment, the extrinsic calibration is carried out using 100 2D feature points and 100 3D feature points. Fig. 8 shows the mean error distribution for computed poses of all cameras, for fixed 3D noise and varying 2D noise. Fig. 9 shows the mean distribution in case the 2D noise is fixed. These results again show the improved performance of the proposed approaches due to the use of both 2D and 3D information together, with the help of correct weighting factor. The performance of all methods gets affected as the noise in 2D and 3D data increases. These results also show improvement in performance of all methods as compared to the multi-view system composed of two cameras due to increased number of 2D and 3D points



(a) Mean Rotation Error

(b) Mean Relative Translation Error

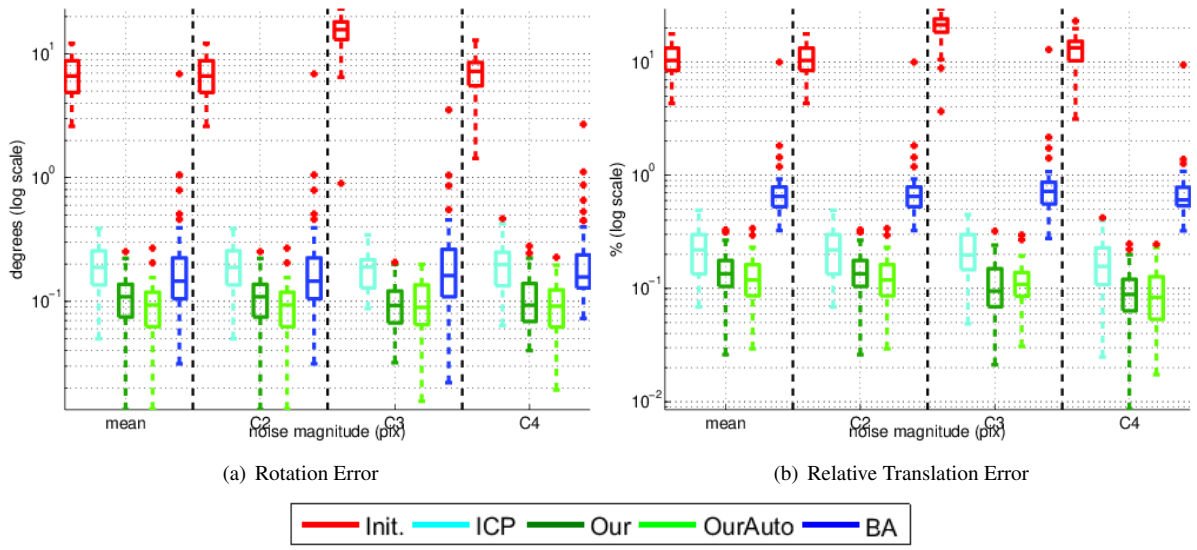
Fig. 9. *Mean Error distribution, of pose estimates for cameras **C2**, **C3** and **C4**. in a four camera setup. 100 2D and 100 3D feature points are used. Gaussian noise is added to the data, being the variance of the 2D noise fixed ($\sigma_{2D} = 1pix$), and the 3D noise σ_{3D} is varied between $8mm$ and $30mm$ (horizontal axes).*

available. Moreover, these results show that the proposed scheme generalizes for different numbers of cameras used in the multi-view system.

We also notice an interesting behavior where in some cases the proposed automatic iterative scheme based on alternative computation of camera poses and w gives better results compared to the scheme with known w. Apart from increase in the number of measurements per feature point, a reason for this can be that for the case of known w we are assuming that for all the 2D and 3D feature points the variances of noise affecting them are the same and constant; but depending on a particular realization, the noise will be a bit higher or lower than the fixed value. Therefore the automatic procedure which tries to compute the variances directly from the noisy data is, in many cases, better able to capture the noise characteristics. For BA, Fig. 9 shows a decrease in its performance as the 3D noise increases. The reason being that apart from its dependence on the initial camera poses, the initial guess of the 3D points corresponding to 2D feature points also gets worse due to increased 3D noise.

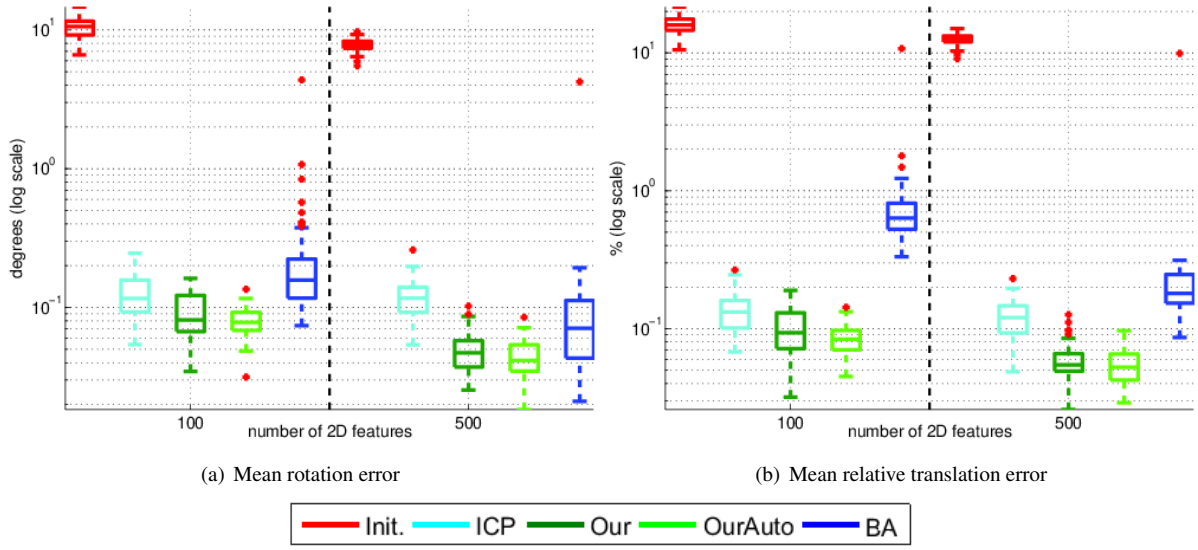
In Fig. 10, we compare the mean error distribution with error distributions of individual cameras for the single case of 2D and 3D noise variance ($\sigma_{2D} = 1$pix, $\sigma_{3D} = 18$mm). These results show that while the initial guess for camera ***C3*** is comparatively worse, the performance of optimization schemes is comparable across all views.

7.3.2. Varying Number of Points: In this experiment, the extrinsic calibration is carried out using a fixed noise variance ($\sigma_{2D} = 1$pix, $\sigma_{3D} = 18$mm). Fig. 11 shows the mean error distribution for a fixed number of 3D points and a varying number of 2D points. Fig. 12, on the other hand, shows the mean distribution in case the 2D points are kept fixed, and the number of 3D points are varied. Here, again, the conclusions drawn in the previous sections hold, while also showing that increasing the number of data points of one of the sensor modalities always improves the extrinsic calibration accuracy for the methods using those modalities.



(a) Rotation Error (b) Relative Translation Error

Fig. 10. *Comparison of error distributions, of the extrinsic calibration of a four camera setup, using 100 2D and 100 3D feature points. The results are based on mean error distribution and error distribution for camera* ***C2****, camera* ***C3*** *and camera* ***C4****. Gaussian noise is added to the data, being the variance of the both 2D noise and 3D noise fixed ($\sigma_{2D} = 1pix$, $\sigma_{3D} = 18mm$).*

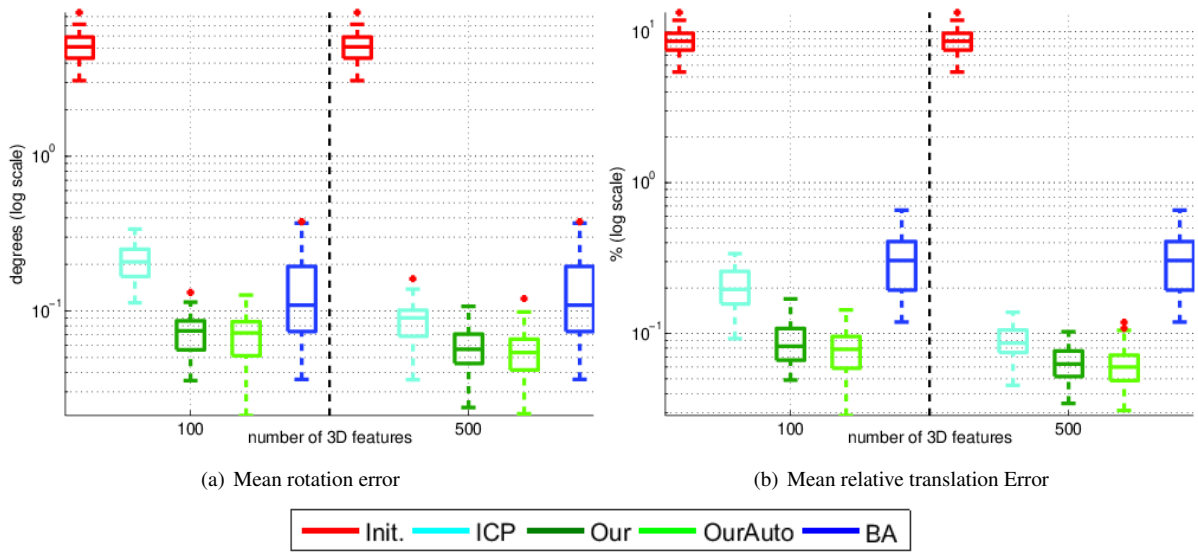


(a) Mean rotation error (b) Mean relative translation error

Fig. 11. *Mean error distribution, of pose estimates for cameras* ***C2****,* ***C3*** *and* ***C4****. in a four camera setup. Gaussian noise is added to the data ($\sigma_{2D} = 1pix$), $\sigma_{3D} = 18mm$), 250 3D feature points and a varying number of 2D feature points (horizontal axes) is considered.*

8. Experiments with Real Data

In this section, we carry out a qualitative performance analysis of the proposed bi-objective refinement scheme using a real setup. Our setup consists of 4 Asus Xtion Pro Live cameras [2] with their positions shown in Fig. 13. Each camera acquires an RGB image and a depth image which



(a) Mean rotation error

(b) Mean relative translation Error

Fig. 12. *Mean error distribution, of pose estimates for cameras **C2**, **C3** and **C4**. in a four camera setup. Gaussian noise is added to the data ($\sigma_{2D} = 1pix$), $\sigma_{3D} = 18mm$), 250 2D feature points and a varying number of 3D feature points (horizontal axes) is considered.*

is mapped to the RGB image. The first step is to perform intrinsic calibration to find the intrinsic and distortion parameters for each camera. For this purpose, we use the method proposed by Zhang [40] which uses 2D corners extracted from RGB images of a checkerboard pattern viewed at different poses to compute these parameters [10]. As mentioned before, the measurements of these RGB-D cameras suffer from inherent depth bias. Therefore, we perform a depth bias correction procedure, similar to the one used in [25], for each camera separately. This procedure requires placing the camera at known distances away from an object (a plane in our case). Using known and measured depth values, we estimate the coefficients of a polynomial which computes the depth correction as a function of measured depth value. These coefficients are unique to each camera and, hence, are used to correct the depth measurements acquired by that camera.

After intrinsic calibration and depth bias correction, the next step is to perform the extrinsic calibration using the proposed bi-objective scheme. We first need to extract matching 2D and 3D feature points using RGB and depth images acquired by all 4 cameras. We again use different views of a (two-sided) planar checkerboard pattern as shown in Fig. 13 and extract matching corners from RGB images to be used as 2D feature points and use the corresponding depth values from depth images to get the 3D feature points. The 3D feature points are filtered via a plane detection approach based on RANSAC algorithm to remove outliers if any exist. For this experiment, only 59 2D and 3D feature points were used. The initial pose estimates are generated in the same manner as explained in Section 7, via a Direct Linear Transform (DLT) based approach [10, 16]. These initial poses are then refined via the proposed iterative pose estimation and weight estimation approach explained in Section 6, BA and ICP. Once the refined poses are obtained, they can be used to produce full, textured, 3D reconstructions using data acquired by all 4 cameras as shown in Fig. 14. A qualitative comparison of 3D reconstructions obtained via different calibration methods is shown in Fig. 15. It can be seen that the partial reconstructions are better aligned using the proposed method, which means that the quality of the extrinsic calibration is superior when compared to the other approaches. Note that we are only showing the alignment of the partial point clouds, and no

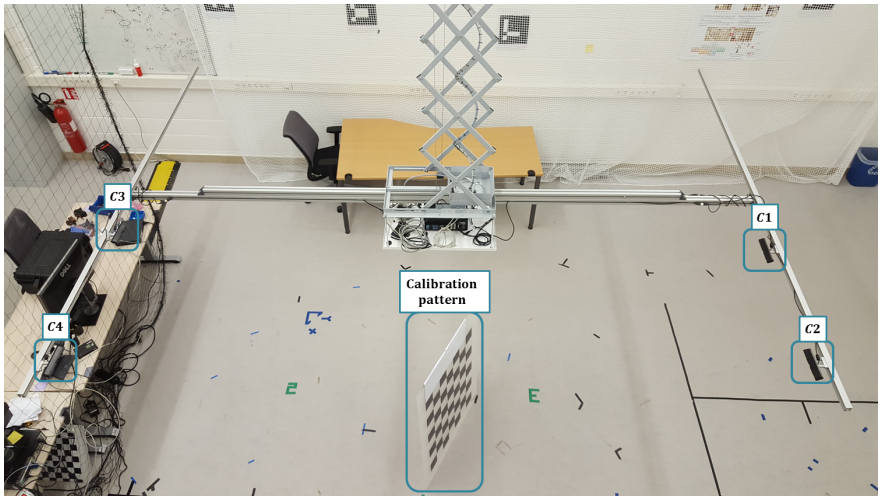


Fig. 13. *Multi-view system consisting of 4 Asus Xtion Pro Live Cameras* ***C1****,* ***C2****,* ***C3*** *and* ***C4*** *mounted on a ceiling lift. This system is used to acquire measurements of a real scene. A two-sided planar checkerboard calibration pattern used to extract feature points is also shown.*

post-processing step such as smoothing or meshing are applied. We chose to do so to better assess, visually, the accuracy of the extrinsic calibration.

9. Conclusion

In this paper we have proposed a framework for RGB and depth sensor fusion based on bi-objective optimization, for refinement of extrinsic calibration in RGB-D multi-view systems. Our bi-objective optimization scheme makes use of a cost function from the BA algorithm for 2D feature points extracted from RGB images and a cost function from the ICP algorithm for 3D feature points extracted from depth images. We analytically derive an expression for the weighted bi-objective cost function. It also analytically relates the weighing factor to the noise in the 2D and 3D measurements, thus making the cost function free of any parameter that needs to be tuned. In case the information regarding measurement noise in 2D and 3D data is not available, we propose an iterative scheme which alternates between estimation of noise parameters assuming known poses, and estimation of camera poses assuming known noise parameters. Thus, it enables us to automatically compute the correct weighting factor when information about measurement noise is not available. A thorough investigation of the performance of the proposed approach for both synthetic and real data showed improved accuracy compared to refinement schemes which only use 2D or 3D information, and comparative performance of proposed approaches with known and unknown noise parameters. These experiments also showed the invariance of the proposed approach under various conditions which include varying the number and position of cameras, varying the 2D and 3D noise and varying the number of the 2D and 3D feature points.

10. Acknowledgment

This work was supported by the National Research Fund (FNR), Luxembourg, under the CORE project C11/BM/1204105/FAVE/Ottersten.

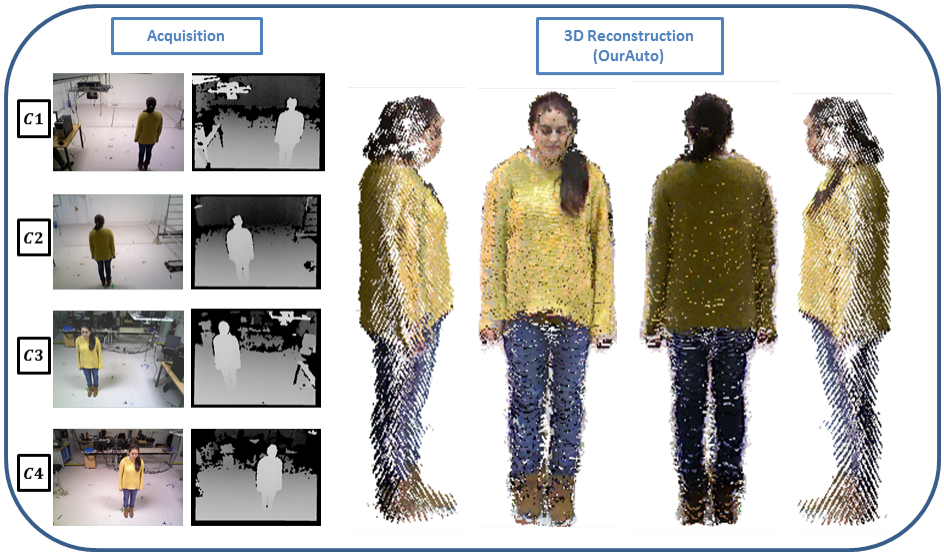


Fig. 14. *3D reconstruction of a human using a real scene acquired from the multi-view system shown in Fig. 13. Acquisition: Each of the 4 cameras acquire an RGB image and a depth image. 3D Reconstruction: Point clouds based 3D reconstruction using pose estimates refined by the proposed bi-objective scheme with the help of automated weighting.*

References

[1] V-REP http://www.coppeliarobotics.com/. URL http://www.primesense.com/

[2] Xtion PRO LIVE https://www.asus.com/3D-Sensor/Xtion_PRO_LIVE/. URL https://www.asus.com/3D-Sensor/Xtion_PRO_LIVE/

[3] Afzal, H., Aouada, D., Fofi, D., Mirbach, B., Ottersten, B.: RGB-D Multi-view System Calibration for Full 3D Scene Reconstruction. In: Pattern Recognition (ICPR), 2014 22nd International Conference on, pp. 2459–2464 (2014). DOI 10.1109/ICPR.2014.425

[4] Alexiadis, D., Kordelas, G., Apostolakis, K., Agapito, J., Vegas, J., Izquierdo, E., Daras, P.: Reconstruction for 3D immersive virtual environments. In: Image Analysis for Multimedia Interactive Services (WIAMIS), 2012 13th International Workshop on, pp. 1–4 (2012). DOI 10.1109/WIAMIS.2012.6226760

[5] Amplianitis, K., Adduci, M., Reulke, R.: Calibration of a Multiple Stereo and Rgb-D Camera System for 3d Human Tracking. ISPRS - International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences (1), 7–14 (2014). DOI 10.5194/isprsarchives-XL-3-W1-7-2014

[6] Berger, K.: A State of the Art Report on Multiple RGB-D Sensor Research and on Publicly Available RGB-D Datasets. In: L. Shao, J. Han, P. Kohli, Z. Zhang (eds.) Computer Vision and Machine Learning with RGB-D Sensors, Advances in Computer Vision and Pattern Recognition, pp. 27–44. Springer International Publishing (2014). DOI 10.1007/978-3-319-08651-4_2. URL http://dx.doi.org/10.1007/978-3-319-08651-4_2

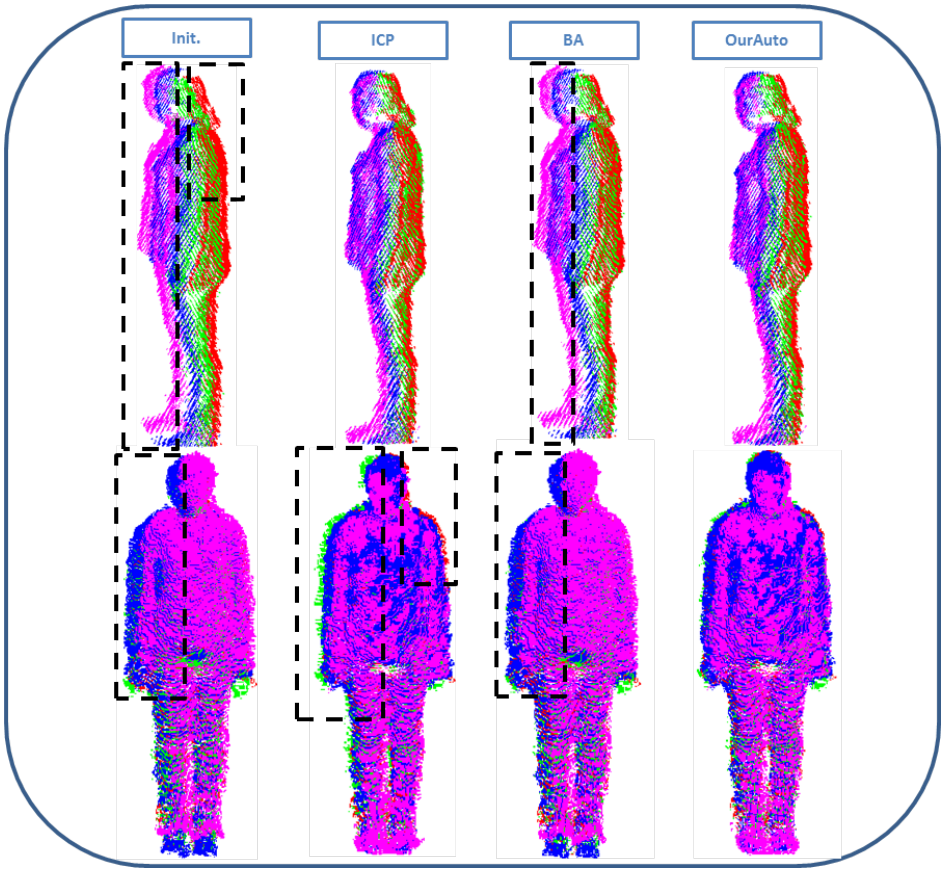


Fig. 15. *Comparison of 3D reconstructions of a human using a real scene as shown in Fig. 13, via different calibration methods namely Init., ICP, BA and OurAuto described in Fig. 5. The acquisitions from cameras **C1**, **C2**, **C3** and **C4** are assigned the colors red, green, blue and magenta, respectively. Misalignments are highlighted via black boxes. Top Row shows side view of the 3D reconstruction and misalignment of views in the results of Init. and BA can be seen clearly, while Bottom Row shows the frontal view and misalignment of views in the results of Init., ICP and BA are visible. It can also be seen that OurAuto gives better results compared to the other methods.*

[7] Berger, K., Ruhl, K., Albers, M., Schroder, Y., Scholz, A., Kokemuller, J., Guthe, S., Magnor, M.: The capturing of turbulent gas flows using multiple Kinects. In: Computer Vision Workshops (ICCV Workshops), 2011 IEEE International Conference on, pp. 1108–1113 (2011). DOI 10.1109/ICCVW.2011.6130374

[8] Berger, K., Ruhl, K., Brümmer, C., Schröder, Y., Scholz, A., Magnor, M.: Markerless Motion Capture using multiple Color-Depth Sensors. In: Proc. Vision, Modeling and Visualization (VMV) 2011, pp. 317–324 (2011)

[9] Besl, P.J., McKay, N.D.: A method for registration of 3-d shapes. IEEE Trans. Pattern Anal. Mach. Intell. **14**(2)

[10] Bougouet, J.: `http://www.vision.caltech.edu/bouguetj/calib doc/` (2007)

[11] Dou, M., Fuchs, H.: Temporally enhanced 3D capture of room-sized dynamic scenes with

commodity depth cameras. In: Virtual Reality (VR), 2014 iEEE, pp. 39–44 (2014). DOI 10.1109/VR.2014.6802048

[12] Dou, M., Guan, L., Frahm, J., Fuchs, H.: Exploring High-Level Plane Primitives for Indoor 3D Reconstruction with a Hand-held RGB-D Camera. In: Computer Vision - ACCV 2012 Workshops, ACCV 2012 International Workshops, Daejeon, Korea, November 5-6, 2012, Revised Selected Papers, Part II, pp. 94–108 (2012). DOI 10.1007/978-3-642-37484-5_9. URL http://dx.doi.org/10.1007/978-3-642-37484-5_9

[13] Dou, M., Taylor, J., Fuchs, H., Fitzgibbon, A., Izadi, S.: 3D Scanning Deformable Objects With a Single RGBD Sensor (2015)

[14] Fankhauser, P., Bloesch, M., Rodriguez, D., , Kaestner, R., Hutter, M., Siegwart, R.: Kinect v2 for Mobile Robot Navigation: Evaluation and Modeling. In: IEEE International Conference on Advanced Robotics (ICAR) (2015)

[15] Han, T., Xu, C., Loxton, R., Xie, L.: Bi-objective optimization for robust RGB-D visual odometry. In: Control and Decision Conference (CCDC), 2015 27th Chinese, pp. 1837–1844 (2015). DOI 10.1109/CCDC.2015.7162218

[16] Hartley, R.I., Zisserman, A.: Multiple View Geometry in Computer Vision, second edn. Cambridge University Press, ISBN: 0521540518 (2004)

[17] Henry, P., Krainin, M., Herbst, E., Ren, X., Fox, D.: RGB-D mapping: Using kinect-style depth cameras for dense 3D modeling of indoor environments. International Journal of Robotics Research (IJRR) **31**(5), 647–663 (2012)

[18] Kainz, B., Hauswiesner, S., Reitmayr, G., Steinberger, M., Grasset, R., Gruber, L., Veas, E., Kalkofen, D., Seichter, H., Schmalstieg, D.: OmniKinect: real-time dense volumetric data acquisition and applications. In: Proceedings of the 18th ACM symposium on Virtual reality software and technology, VRST '12, pp. 25–32. ACM, New York, NY, USA (2012). DOI 10.1145/2407336.2407342. URL http://doi.acm.org/10.1145/2407336.2407342

[19] Kuster, C., Popa, T., Zach, C., Gotsman, C., Gross, M.: FreeCam: A Hybrid Camera System for Interactive Free-Viewpoint Video. In: Proceedings of Vision, Modeling, and Visualization (VMV) (2011)

[20] Maimone, A., Fuchs, H.: Encumbrance-free telepresence system with real-time 3D capture and display using commodity depth cameras. In: Mixed and Augmented Reality (ISMAR), 2011 10th IEEE International Symposium on, pp. 137–146 (Oct.). DOI 10.1109/ISMAR.2011.6092379

[21] Michot, J., Bartoli, A., Gaspard, F.: Bi-Objective Bundle Adjustment With Application to Multi-Sensor SLAM. In: 3DPVT'10 – Int'l Symp. on 3D Data Processing, Visualization and Transmission. Paris, France (2010)

[22] Miller, S., Teichman, A., Thrun, S.: Unsupervised extrinsic calibration of depth sensors in dynamic scenes. In: Intelligent Robots and Systems (IROS), 2013 IEEE/RSJ International Conference on, pp. 2695–2702 (2013). DOI 10.1109/IROS.2013.6696737

[23] Moré, J.: The Levenberg-Marquardt algorithm: Implementation and theory. In: G.A. Watson (ed.) Numerical Analysis, *Lecture Notes in Mathematics*, vol. 630, chap. 10, pp. 105–116–116. Springer Berlin / Heidelberg (1978). DOI 10.1007/bfb0067700. URL http://dx.doi.org/10.1007/bfb0067700

[24] Myung, I.J.: Tutorial on maximum likelihood estimation. Journal of Mathematical Psychology **47**(1), 90–100 (2003)

[25] Nakazawa, M., Mitsugami, I., Makihara, Y., Nakajima, H., Habe, H., Yamazoe, H., Yagi, Y.: Dynamic scene reconstruction using asynchronous multiple Kinects. In: Pattern Recognition (ICPR), 2012 21st International Conference on, pp. 469–472 (2012)

[26] Newcombe, R.A., Izadi, S., Hilliges, O., Molyneaux, D., Kim, D., Davison, A.J., Kohli, P., Shotton, J., Hodges, S., Fitzgibbon, A.: KinectFusion: Real-time Dense Surface Mapping and Tracking. In: Proceedings of the 2011 10th IEEE International Symposium on Mixed and Augmented Reality, ISMAR '11, pp. 127–136. IEEE Computer Society, Washington, DC, USA (2011). DOI 10.1109/ISMAR.2011.6092378. URL http://dx.doi.org/10.1109/ISMAR.2011.6092378

[27] Nüchter, A., Elseberg, J., Schneider, P., Paulus, D.: Study of parameterizations for the rigid body transformations of the scan registration problem. Computer Vision and Image Understanding **114**(8), 963–980 (2010)

[28] Palasek, P., Yang, H., Xu, Z., Hajimirza, N., Izquierdo, E., Patras, I.: A flexible calibration method of multiple Kinects for 3D human reconstruction. In: Multimedia Expo Workshops (ICMEW), 2015 IEEE International Conference on, pp. 1–4 (2015). DOI 10.1109/ICMEW.2015.7169829

[29] Penelle, B., Schenkel, A., Warzee, N.: Geometrical 3D reconstruction using real-time RGB-D cameras. In: 3D Imaging (IC3D), 2011 International Conference on, pp. 1–8 (2011). DOI 10.1109/IC3D.2011.6584368

[30] Prince, S.J.D.: Computer Vision: Models, Learning, and Inference, 1st edn. Cambridge University Press, New York, NY, USA (2012)

[31] Rabaud, V.: Vincent's Structure from Motion Toolbox. http://github.com/vrabaud/sfm_toolbox

[32] Segal, A., Haehnel, D., Thrun, S.: Generalized-ICP. In: Proceedings of Robotics: Science and Systems. Seattle, USA (2009)

[33] Smisek, J., Jancosek, M., Pajdla, T.: 3D with Kinect. In: ICCV Workshops, pp. 1154–1160. IEEE (2011). URL http://dblp.uni-trier.de/db/conf/iccvw/iccvw2011.html#SmisekJP11

[34] Sturm, J., Bylow, E., Kahl, F., Cremers, D.: CopyMe3D: Scanning and printing persons in 3D. In: German Conference on Pattern Recognition (GCPR). Saarbrücken, Germany (2013)

[35] Teng, D., Bazin, J.C., Martin, T., Kuster, C., Cai, J., Popa, T., Gross, M.: Registration of Multiple RGBD Cameras via Local Rigid Transformations. IEEE International Conference on Multimedia & Expo (2014)

[36] Triggs, B., McLauchlan, P., Hartley, R., Fitzgibbon, A.: Bundle Adjustment – A Modern Synthesis. In: VISION ALGORITHMS: THEORY AND PRACTICE, LNCS, pp. 298–375. Springer Verlag (2000)

[37] Tykkala, T., Audras, C., Comport, A.: Direct Iterative Closest Point for real-time visual odometry. In: Computer Vision Workshops (ICCV Workshops), 2011 IEEE International Conference on, pp. 2050–2056 (2011). DOI 10.1109/ICCVW.2011.6130500

[38] Yang, R., Chan, Y.H., Gong, R., Nguyen, M., Strozzi, A., Delmas, P., Gimel'farb, G., Ababou, R.: Multi-Kinect scene reconstruction: Calibration and depth inconsistencies. In: Image and Vision Computing New Zealand (IVCNZ), 2013 28th International Conference of, pp. 47–52 (2013). DOI 10.1109/IVCNZ.2013.6726991

[39] Zhang, Q.: Extrinsic calibration of a camera and laser range finder. In: In IEEE International Conference on Intelligent Robots and Systems (IROS, p. 2004 (2004)

[40] Zhang, Z.: Flexible camera calibration by viewing a plane from unknown orientations. In: Computer Vision, 1999. The Proceedings of the Seventh IEEE International Conference on, vol. 1, pp. 666–673 vol.1 (1999). DOI 10.1109/ICCV.1999.791289

11. Appendix

11.1. Non-linear Optimization for Proposed Bi-Objective Framework

Due to the non-linear dependence of cost function in (13) on parameters in $\mathbf{S}$, the MLE $\hat{\mathbf{S}}$ is to be computed via a numerical scheme based on non-linear optimization. In this scheme at every iteration a small change is introduced in the current set of parameters leading to comparatively improved performance or lower residual [24]. First step in this scheme is to linearize $\mathbf{b}^j_{l,m}(\mathbf{T}_l, \mathbf{T}_m)$ and $\mathbf{a}^h_l(\mathbf{S}^h_l)$ about current estimate $\hat{\mathbf{S}}$ assuming very small error $\Delta\mathbf{S}$ using Taylor expansion to get:

$$\mathbf{b}^j_{l,m}(\mathbf{T}_l, \mathbf{T}_m) \approx \mathbf{b}^j_{l,m}(\hat{\mathbf{T}}_l, \hat{\mathbf{T}}_m) + \mathbf{J}_{\mathbf{b}^j_{l,m}} \Delta\mathbf{S}, \tag{16}$$

and:

$$\mathbf{a}^h_l(\mathbf{S}^h_l) \approx \mathbf{a}^h_l(\hat{\mathbf{S}}^h_l) + \mathbf{J}_{\mathbf{a}^h_l} \Delta\mathbf{S}, \tag{17}$$

where $\mathbf{J}_{\mathbf{b}^j_{l,m}}$ and $\mathbf{J}_{\mathbf{a}^h_l}$ are Jacobians of $\mathbf{b}^j_{l,m}(\mathbf{T}_l, \mathbf{T}_m)$ and $\mathbf{a}^h_l(\mathbf{S}^h_l)$, with respect to $\mathbf{S}$, respectively. Replacing (16) and (17) in (13) and concatenating $\mathbf{b}^j_{l,m}$, $\mathbf{a}^h_l$, $\Delta\mathbf{S}$ and corresponding Jacobians we have:

$$\begin{aligned} V(\mathbf{S}) &\approx (\mathbf{B} + \mathbf{J}_\mathbf{B}\boldsymbol{\Delta}\mathbf{S})^T(\mathbf{B} + \mathbf{J}_\mathbf{B}\boldsymbol{\Delta}\mathbf{S}) + \\ &\quad w(\mathbf{A} + \mathbf{J}_\mathbf{A}\boldsymbol{\Delta}\mathbf{S})^T(\mathbf{A} + \mathbf{J}_\mathbf{A}\boldsymbol{\Delta}\mathbf{S}) \\ &= (\mathbf{B}^T\mathbf{B} + 2\Delta\mathbf{S}^T\mathbf{J}^T_\mathbf{B}\mathbf{B} + \Delta\mathbf{S}^T\mathbf{J}^T_\mathbf{B}\mathbf{J}_\mathbf{B}\Delta\mathbf{S}) \\ &\quad + w(\mathbf{A}^T\mathbf{A} + 2\Delta\mathbf{S}^T\mathbf{J}^T_\mathbf{A}\mathbf{A} + \Delta\mathbf{S}^T\mathbf{J}^T_\mathbf{A}\mathbf{J}_\mathbf{A}\Delta\mathbf{S}). \end{aligned} \tag{18}$$

After that we take the derivative of $V(\mathbf{S})$ with respect to $\mathbf{S}$ and equate it to zero to get:

$$\frac{\partial V(\mathbf{S})}{\partial \mathbf{S}} \approx \mathbf{J}^T_\mathbf{B}\mathbf{B} + \mathbf{J}^T_\mathbf{B}\mathbf{J}_\mathbf{B}\boldsymbol{\Delta}\mathbf{S} + w\mathbf{J}^T_\mathbf{A}\mathbf{A} + w\mathbf{J}^T_\mathbf{A}\mathbf{J}_\mathbf{A}\boldsymbol{\Delta}\mathbf{S} = 0. \tag{19}$$

Rearranging (19), we get the parameter update rule as:

$$\mathbf{\Delta S} = -(\mathbf{J}_\mathbf{B}^T\mathbf{J}_\mathbf{B} + w\mathbf{J}_\mathbf{A}^T\mathbf{J}_\mathbf{A})^{-1}(\mathbf{J}_\mathbf{B}^T\mathbf{B} + w\mathbf{J}_\mathbf{A}^T\mathbf{A}). \tag{20}$$

We can also rearrange (20) according to Levenberg-Marquardt LM [23] algorithm get the parameter update rule as:

$$\begin{aligned}((\frac{1}{2\sigma_{3D}^2}\mathbf{J}_\mathbf{B}^T\mathbf{J}_\mathbf{B} + \frac{1}{\sigma_{2D}^2}\mathbf{J}_\mathbf{A}^T\mathbf{J}_\mathbf{A}) + \\ \lambda\texttt{diag}(\frac{1}{2\sigma_{3D}^2}\mathbf{J}_\mathbf{B}^T\mathbf{J}_\mathbf{B} + \frac{1}{\sigma_{2D}^2}\mathbf{J}_\mathbf{A}^T\mathbf{J}_\mathbf{A}))\mathbf{\Delta S} \\ = -(\frac{1}{2\sigma_{3D}^2}\mathbf{J}_\mathbf{B}^T\mathbf{B} + \frac{1}{\sigma_{2D}^2}\mathbf{J}_\mathbf{A}^T\mathbf{A}),\end{aligned} \tag{21}$$

where λ is the damping factor.