Error Accuracy Estimation of 3D Reconstruction and 3D Camera Pose from RGB-D Data

Luis E. Ortiz-Fernandez
Electr. and Comp. Eng. Grad. Program
Federal Univ. of Rio Grande do Norte
Natal, Brazil
ORCID: 0000-0002-8578-4515

Bruno M. F. Silva
Mechatronic Eng. Grad. Program
Federal Univ. of Rio Grande do Norte
Natal, Brazil
ORCID: 0000-0002-7780-7254

Luiz M. G. Gonçalves
Electr. and Comp. Eng. Grad. Program
Federal Univ. of Rio Grande do Norte
Natal, Brazil
ORCID: 0000-0002-7735-5630

***Abstract*—We propose an approach to predict accuracy for three-dimensional reconstruction and camera pose using a generic RGB-D camera on a robotic platform. We initially create a ground truth of 3D points and camera poses using a set of smart markers that we specifically devised and constructed for our approach. Then, we compute actual errors and their accuracy during the motion of our mobile robotic platform. A modeling of the error is then provided, which is used as input to a deep multi-layer perceptron in order to estimate accuracy as a function of the camera's distance, velocity, and vibration of the vision system. The network outputs are the root mean squared errors for the 3D reconstruction and the relative pose errors for the camera. Experimental results show that this approach has a prediction accuracy of $\pm 1\%$ for the 3D reconstruction and $\pm 2.5\%$ for camera poses, which shows a better performance in comparison with state-of-the-art methods.**

***Index Terms*—Errors Prediction, Camera Positioning, 3D Reconstruction, RGB-D Cameras**

I. Introduction

Two known problems that appear in everyday vision tasks are the scene 3D reconstruction [1]–[3] and the estimation of camera pose [4], [5]. These problems are generally tackled by using RGB-D devices with technologies as Structured Light (SL), e.g., MS Kinect v1, Time of Light (ToF), e.g., MS Kinect v2, or Coded Light (CL), e.g, Intel RealSense^{TM} Stereo depth technology, or even with purely stereoscopic vision devices as the Minoru [6], ZED from Stereolabs [7] or BumbleBee from Point Gray [8]. Determination of position and orientation from monocular cameras has also been done in the literature, mainly for simultaneous localization and mapping (SLAM) using robots [9]. Despite being very useful, these low-cost RGB-D cameras are not highly accurate due to various internal and external factors that may cause errors in their measurements. Internal factors cause systematic errors that depend on the camera hardware, such as lenses with high distortion, an inadequate arrangement of cameras, and the use of sensors with low resolution. The external factors are those that affect the camera's performance independent of its construction characteristics, causing random errors. Some external factors are velocity, vibrations, lighting, and others. In this work we are most interested in providing a way to measuring, modeling, and estimating the accuracy of the error caused by these external factors.

This work is partially financed by CNPq and CAPES under grant 001 - Brazil.

From the methods found in the literature to quantify errors in RGB-D camera measurements, there are a few works that include the analysis of the systematic errors in stereo 3D reconstruction in situations where the camera is static [7], [10]. On the same direction of this paper, there are works that model the RGB-D measurement errors due to the effects of vibration camera internal factors [11], [12]. Nevertheless, these works and other works that will be mentioned further in Section III do not make a joint analysis of the effects that camera displacement, velocity, and vibrations have on the measurements.

Hence, errors caused by external factors are rarely considered in robotic vision applications because of the lack of a single method to quantify it, mainly due to the limited literature. Thus, in this paper, we propose a methodology that includes a complete pipeline with an easy of implementation technique to measure, model, and predict errors accuracy in 3D camera pose and 3D reconstruction using distance, velocity, and vibration as input. Our contribution is a novel approach in regard to the literature, providing a useful way to estimate efficacy of methods and devices used in tasks that here are basis for robotic vision. As well, there are applications as virtual reality and others involving visual mapping that need motion and position estimation from RGB-D cameras.

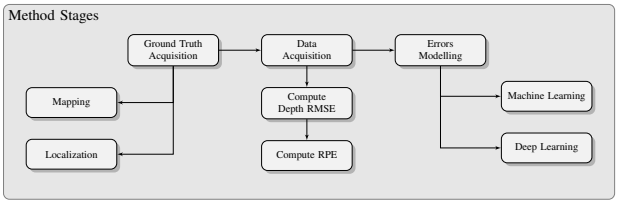


Fig. 1: Stages of the proposed method to measure, model and predict the 3D reconstruction and camera pose errors.

II. The Basic Methodology

We follow a general methodology for measuring, modeling, and predicting error accuracy in 3D reconstruction and camera pose that comprises three stages as shown in Figure 1. Basically, we start with the creation of the ground truth, which

is implemented in two sub-stages, *mapping* to get accurate 3D points (commonly called point cloud) and *localization* to get actual camera poses from visual odometry. The *mapping* and *localization* are normally implemented using a graph optimization method. To compute the error (or difference) between two point clouds, a common approach is to apply a registration method [13], [14]. Usually, the result is a rigid transformation $\mathbf{T}$ that is composed by a rotation matrix $\mathbf{R}$ and a translation vector $\mathbf{t}$ that align the source cloud in the reference system of the target cloud. The rigid transformation $\mathbf{T}$ can be computed by using SVD [15], quaternions [16], or dual quaternions [17].

The second stage of our approach is the acquisition of an errors dataset using an RGB-D camera and a mobile robot. In this paper we analyze the camera random errors caused by vibrations, distance, and velocity, and quantify its accuracy using the root mean square error (RMSE) metric for 3D reconstruction errors [10] and the relative pose error (RPE) metric for camera pose errors [18].

Finally, in the third stage, recorded datasets are used to evaluate a multi-Layer perceptrons (MLP) algorithm to model and predict the 3D reconstruction and camera pose errors in function of the camera's distance, velocity, and vibration.

Thus, as stated above, the core of the problems is how to determine the RMSE in 3D reconstruction and also the (RGB-D) camera RPE, seen next.

A. Three-Dimensional Reconstruction Error

Here, the problem is to get the three-dimensional reconstruction RMSE in the frame $\hat{\mathrm{f}}$ given that an RGB-D camera is moving at velocity $\tilde{\mathbf{s}}$, with vibration $\tilde{\mathbf{v}}$, at a distance $\tilde{\mathbf{z}}$ from origin position $G(0,0,0)$. Also, to determine if it is possible obtaining two sets of points referenced to the same coordinate system, one measured as $\hat{\wp}_{\mathrm{i}}$ (from artificial markers corners detection), and another estimated as $\wp_{\mathrm{i}}$ (ground truth). Consider that in frame $\hat{\mathrm{f}}$ one point $\hat{\mathbf{p}}$ is visualized (with 3D spatial errors due to systematic and random errors). Also consider that the world coordinates $(\hat{\mathrm{X}}, \hat{\mathrm{Y}}, \hat{\mathrm{Z}})$ of the point are known. That is, $\mathbf{p}$ is an error-free version of $\hat{\mathbf{p}}$, with coordinates in the world $(\mathrm{X}, \mathrm{Y}, \mathrm{Z})$. As such, if $\mathrm{i} = 1, .., \mathrm{n}$ points are detected in the image $\hat{\mathrm{f}}$, for a given point $\hat{\mathbf{p}}_i$, the magnitude of its location error at the reference frame can be simply defined by the Euclidean distance e_{i} between points $\mathbf{p}_i$ and $\hat{\mathbf{p}}_{\mathbf{i}}$ as:

$$e_{\mathrm{i}} = \left(e_{i,x}^2 + e_{i,y}^2 + e_{i,z}^2\right)^{1/2} \tag{1}$$

where $e_{i,x} = \mathrm{X_i} - \hat{\mathrm{X}}_{\mathrm{i}}$, $e_{i,y} = \mathrm{Y_i} - \hat{\mathrm{Y}}_{\mathrm{i}}$, and $e_{i,z} = \mathrm{Z_i} - \hat{\mathrm{Z}}_{\mathrm{i}}$.

For all of the detected points the analysis mentioned above can be extended to use the points sets, $\hat{\wp}_{\hat{\mathrm{f}}} = [\hat{\mathbf{p}}_1, ..., \hat{\mathbf{p}}_{\mathrm{n}}]$ and $\wp_{\hat{\mathrm{f}}} = [\mathbf{p}_1, ..., \mathbf{p}_{\mathrm{n}}]$. Hence, the spatial localization errors for the points in the frame $\hat{\mathrm{f}}$ can be computed using the root mean square error (RMSE) as given by Eq. 2 [10].

$$\begin{aligned} RMSE_{\hat{\mathrm{f}},x} &= \left(\tfrac{1}{\mathrm{n}} \textstyle\sum_{\mathrm{i}=1}^{\mathrm{n}} e_{\mathrm{i},x}^2\right)^{1/2} \\ RMSE_{\hat{\mathrm{f}},y} &= \left(\tfrac{1}{\mathrm{n}} \textstyle\sum_{\mathrm{i}=1}^{\mathrm{n}} e_{\mathrm{i},y}^2\right)^{1/2} \\ RMSE_{\hat{\mathrm{f}},z} &= \left(\tfrac{1}{\mathrm{n}} \textstyle\sum_{\mathrm{i}=1}^{\mathrm{n}} e_{\mathrm{i},z}^2\right)^{1/2} \\ RMSE_{\hat{\mathrm{f}}} &= \left(\tfrac{1}{\mathrm{n}} \textstyle\sum_{\mathrm{i}=1}^{\mathrm{n}} e_{\mathrm{i}}^2\right)^{1/2} \end{aligned} \tag{2}$$

B. Camera Pose Error

Given two camera poses (position and orientation), one measured $\mathbf{C}_{\hat{\mathrm{f}}}$ referenced to a global coordinate system $G(0,0,0)$, and its ground truth $\mathbf{C}_{\hat{\mathrm{f}}}^*$, the problem here is to determine what is the relative error between these two poses if the camera is moving (e.g., within an environment with artificial markers or RTK-GPS) at a distance $\tilde{\mathbf{z}}$ with a velocity $\tilde{\mathbf{s}}$ and with a vibration $\tilde{\mathbf{v}}$. The relative pose error (RPE), denoted here by the matrix $\hat{\mathbf{R}}_{\hat{\mathrm{f}}}$, between the poses $\mathbf{C}_{\hat{\mathrm{f}}}^*$ and $\mathbf{C}_{\hat{\mathrm{f}}}$, can be computed using Eq. 3, where Δ is the interval between the poses (number of frames) [18].

$$\hat{\mathbf{R}}_{\hat{\mathrm{f}}} = \left[\hat{\mathbf{R}}_{\hat{\mathrm{f}},x}, \hat{\mathbf{R}}_{\hat{\mathrm{f}},y}, \hat{\mathbf{R}}_{\hat{\mathrm{f}},z}\right] = \left(\mathbf{C}_{\hat{\mathrm{f}}}^{*-1}\mathbf{C}_{\hat{\mathrm{f}}+\Delta}^{*}\right)^{-1}\left(\mathbf{C}_{\hat{\mathrm{f}}}^{-1}\mathbf{C}_{\hat{\mathrm{f}}+\Delta}\right) \tag{3}$$

If the RPE between two poses $\hat{\mathbf{R}}_{\hat{\mathrm{f}}}$ is computed using only the translation component ($trans$) of $\hat{\mathbf{R}}_{\hat{\mathrm{f}}}$, Eq. 3 is simplified to Eq. 4 [19].

$$\begin{aligned} RPE_{\hat{\mathrm{f}},x} &= \hat{\mathbf{R}}_{\hat{\mathrm{f}},x} \\ RPE_{\hat{\mathrm{f}},y} &= \hat{\mathbf{R}}_{\hat{\mathrm{f}},y} \\ RPE_{\hat{\mathrm{f}},z} &= \hat{\mathbf{R}}_{\hat{\mathrm{f}},z} \\ RPE_{\hat{\mathrm{f}}} &= \|\mathit{trans}(\hat{\mathbf{R}}_{\hat{\mathrm{f}}})\| \end{aligned} \tag{4}$$

If the camera moves within an environment with artificial markers for some time and captures $\hat{\mathrm{f}}$ frames, it is possible to collect two datasets, one of the three-dimensional errors and another of the camera pose errors. Finally, using the collected datasets and regression techniques, it would be possible to obtain the models of Eq. 5 and 6. These models can predict RGB-D measurements errors as a function of distance $\tilde{\mathbf{z}}$, velocity $\tilde{\mathbf{s}}$ and vibration $\tilde{\mathbf{v}}$.

$$\mathcal{N}_{RMSE_{\hat{\mathrm{f}}}} = \mathbf{F}(\tilde{\mathbf{z}}, \tilde{\mathbf{s}}, \tilde{\mathbf{v}}) \tag{5}$$

$$\mathcal{N}_{RPE_{\hat{\mathrm{f}}}} = \mathbf{F}(\tilde{\mathbf{z}}, \tilde{\mathbf{s}}, \tilde{\mathbf{v}}) \tag{6}$$

III. Related Works

We consider two types of methods, when the camera is static and when it is in motion. These methods can be subdivided further into two types of approaches: analytical and experimental. Analytical approaches analyze only the effects of systematic errors due to their mathematical complexity. In contrast, experimental methods allow the analysis of the effects of systematic and random errors.

A. Static Camera

RGB-D cameras can be statically analyzed, for example, when one is expressing the stereo uncertainty as a function of the 3D points' location in the scene [20]. This can be summarized as determining when the stereo uncertainty value reaches the lowest value at the center of the image plane, and decreases as the baseline becomes larger. Another work found addresses the determination of parameters of a stereo system to minimize 3D position errors (in the three axes of the world coordinate system) [21]. Stereo error models considering non-ideal triangulation and an optimal, and finite baseline, are presented. The error propagation from the image coordinate system to the camera system and the world coordinate system is analyzed. Jin et al. [22] show an analytical and numerical analysis of the disparity and 3D measures. The result is a quadratic model for reconstruction error based on the disparity error produced by the radial distortion of the camera lenses.

An experimental work [23] proposes a method for comparison of RGB-D camera resolution and human perception. The authors figured out that the 3D accuracy is more influenced by focal length variations than by a variable baseline. Then, the accuracy of the camera is examined, by analyzing two terms, absolute value and 3D resolution. The investigation of the error due to the distortion parameters, camera calibration of internal and external parameters, and their effects on 3D reconstruction accuracy is also found in the literature [24]. The authors determine that distortion and internal parameters have a minimal impact while the external parameters directly determine 3D reconstruction accuracy by affecting baseline distance and the angle between cameras. In this direction, Sankowski et al. [25] have developed a method to measure the uncertainty of the 3D position of a reconstructed point using an RGB-D camera. The authors show that the total absolute errors of the x and y coordinates of the 3D point have similar values and that these errors are less than the error of the z coordinate.

B. Dynamic Camera

A theoretical analysis of errors of an RGB-D camera mounted on a moving robot [26] shows that the distance error and the header angle increase as the baseline decreases. An experimental work [27] measures the object's 3D location accuracy with an RGB-D camera onboard a boat. Appropriate roll and pitch angles are estimated for eliminating the effects of sudden movements and obtain the relative directions of the camera to evaluate the 3D position of the objects detected in the scene. Actual roll and pitch angles are computed using the horizon data, from which camera rotations are estimated, due to vibration, with two pairs of image coordinates and 3D locations of the landmarks. Okazaki et al. [11] propose an error model for stereo measurements (in pixels and disparity), based on the fact that when the camera vibrates, a point in the stereo pair undergoes displacement in the image coordinates and consequently in the world (on the horizontal axis). The authors determine empirically that the error in the pixel position of a point in the stereo pair has a uniform distribution. The disparity error distribution is of the triangular type. Later on, Okazaki [28] expands the analysis for the horizontal and depth axis.

The representation of error as a function of 3D point motion parameters, such as amplitude, frequency, and phase is also found in the literature [29]. To do this, a calibrated and non-synchronized RGB-D camera takes photos when a spatial point does a simple harmonic motion or sinusoidal movements. Then, the error between the actual and reconstructed z coordinate is computed. The authors determine that the error increases when the amplitude and frequency of the movements are large.

After extensively reviewing the related works, we noticed the inexistence of a work that proposes a method to measure, model, and predict the RGB-D measurements in the function of their distance, velocity, and taking into account vibration problems caused by motion of the robot. Thus, the solution described next is an important contribution.

IV. Measuring, Modeling and Predicting Errors

We need to construct a ground truth for 3D points and camera poses, which is done using a robotic platform coupled with high-precision measurement equipment. The reconstruction and pose errors are measured using this set of actual 3D points.

A. Capturing the Ground Truth Dataset

Most of the methods reported in the literature for creating our ground truth are very costly, and their use is limited to specific workspaces. We use our previous approach that introduces an accurate and low-cost method to create a ground truth using smart markers (SM) [18]. An SM consists of a square planar fiducial marker and a positional measurement unit (PMS). With a set of markers in the environment, the SM $mapping$ and $localization$ implementations return a marker map with actual 3D points (from markers corners) and the camera poses. Previous results demonstrated that the method decreases the RPE by $\approx 85\%$ in the $mapping$ stage and the absolute trajectory error (ATE) in $\approx 50\%$ in the camera $localization$ stage in comparison with other methods [18]. Also, previous results support the claim that it is possible to use low-cost methods for ground-truth generation.

Our data acquisition platform is based on a four-wheel Pioneer 3-AT robot, an onboard computer Jetson AGX Xavier, a Stereolabs ZED camera, a vibration analyzer PCE-VDL-16I, and a laser distance meter GLM80. All devices are calibrated with the measures referenced to the left camera coordinate system, as shown in Figure 2. Communication between the onboard computer and the robot is made using the $Aria$ library [18]. It also allows reading and controlling all sensors to capture RGB-D measurements at variable robot velocities from 0.1 to 0.7 m/sec. Additionally, the onboard computer drives the detection of the marker using the $ArUco$ library.

B. Measuring Three-Dimensional Reconstruction Error

We compute the reconstruction error using a set of actual 3D points obtained from an SM set. Figure 3a shows how the

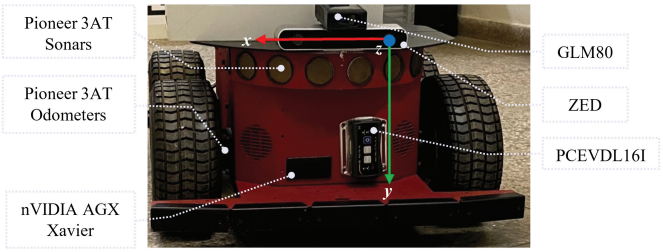
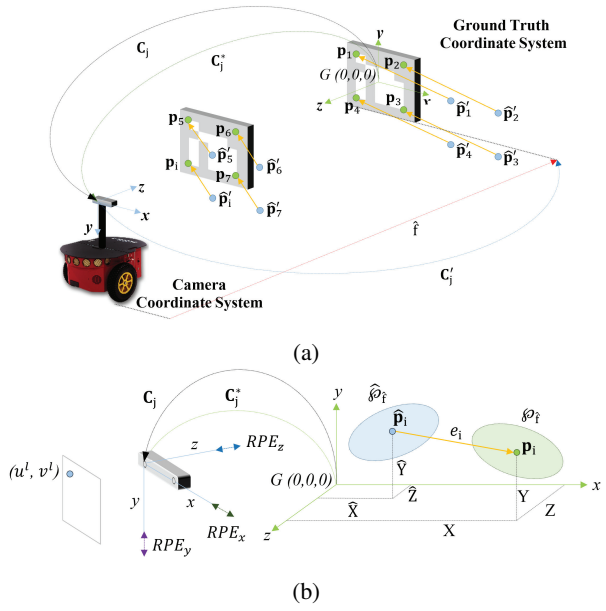


Fig. 2: Data acquisition platform with illustration of the used reference frame (coordinate system). The x–$axis$ is to the right of the robot (red arrow), y is pointing down (green arrow) and z is to the front (forward) of the robot (a right-hand coordinate system is used).



(a)

(b)

Fig. 3: $RMSE_{\hat{f}}$ and $RPE_{\hat{f}}$ computing process. (**a**) Capture actual $\mathbf{p}_i$ (green points) and measured points $\hat{\mathbf{p}}_i^{'}$ (light-blue). Obtain the ground truth camera pose $\mathbf{C}_{\hat{f}}^{*}$ (green dotted arc) and measured pose $\mathbf{C}_{\hat{f}}$ (black dotted arc). (**b**) Computation of error $RMSE_{\hat{f}}$ in all 3D points in the frame $\hat{f}$, using the actual $\wp_{\hat{f}}$ and measured sets. $\hat{\wp}_{\hat{f}}$ is the transformed version of $\hat{\wp}_{\hat{f}}^{'}$. Computation of the $RPE_{\hat{f}}$ and their components.

stereo camera is moving and captures a frame $\hat{f}$ (with the left camera). All visible markers are detected, and the coordinates (u^l, v^l) of their four points are determined using the $ArUco$ library. Then, with these 2D coordinates, the corresponding 3D coordinates for each point $\mathbf{p}_i$ is searched in an optimized marker map.

Thus, in frame $\hat{f}$ the measured 3D points $\hat{\mathbf{p}}_i^{'}$ are obtained using the markers detection and the depth map generated by the stereo camera. All actual 3D points of the set $\wp_{\hat{f}}$ are referenced to $G(0,0,0)$, while all measured points of the set $\hat{\wp}_{\hat{f}}^{'}$ are referenced to the camera coordinate system. We use Eq. 7 to transform all measured points to $G(0,0,0)$. In this equation the camera pose $\mathbf{C}_{\hat{f}}^{*}$ is obtained from the camera $localization$ ground truth data. The transformed measured points set is denoted as $\hat{\wp}_{\hat{f}}$.

$$\hat{\wp}_{\hat{f}} = \mathbf{C}_{\hat{f}}^{*} \hat{\wp}_{\hat{f}}^{'} \tag{7}$$

As shown in Figure 3b, once the actual and measured sets of 3D points are in the same coordinate system, the point error is computed using Eq. 1. Finally all individual errors e_i are put together in the metric $RMSE_{\hat{f}}$ using Eq. 2.

C. Measuring Camera Relative Positioning Error

The camera RPE is the difference between the ground truth pose and the measured pose. In Figure 3a the ground truth for the camera pose $\mathbf{C}_{\hat{f}}^{*}$ is obtained from a camera $localization$ process described in Section IV-A, while that the measured pose $\mathbf{C}_{\hat{f}}^{'}$ is get from ZED SDK. The measured $\mathbf{C}_{\hat{f}}$ is the camera pose in the markers map coordinate system, which is the result of applying the transformation $\mathbf{C}_1^{*}$ to $\mathbf{C}_{\hat{f}}^{'}$ as given by Eq. 8. This transformation is the first camera pose $\mathbf{C}_1^{*}$ given by the ground truth. Applying this transformation is necessary because the measured pose initially is in the camera coordinate system.

$$\mathbf{C}_{\hat{f}} = \mathbf{C}_1^{*} \mathbf{C}_{\hat{f}}^{'} \tag{8}$$

Once the measured and ground truth camera poses are in the same coordinate system, we compute the camera RPE using Eq. 4 with $\Delta = 1$ and only for consecutive poses (not all poses pairs are used).

V. Experimental Evaluation

We recorded data for six trajectories for robot velocities from 0.1 to 0.6 m/s using the data acquisition platform. The scene is a flat hall with marble floor and artificial light. The recorded data are grouped in four datasets: **RMSE-Model/hall**, **RPE-Model/hall**, used for modeling and **RMSE-Eval/hall**, **RPE-Eval/hall** for evaluation. The dataset used for modeling has 90,000 rows while the dataset for evaluation has 67,500 rows. All datasets have twelve columns: the distance ($\tilde{\mathbf{z}}_\mathbf{x}$, $\tilde{\mathbf{z}}_\mathbf{y}$, $\tilde{\mathbf{z}}_\mathbf{z}$), velocity ($\tilde{\mathbf{s}}_\mathbf{x}$, $\tilde{\mathbf{s}}_\mathbf{y}$, $\tilde{\mathbf{s}}_\mathbf{z}$) and vibration ($\tilde{\mathbf{v}}_\mathbf{x}$, $\tilde{\mathbf{v}}_\mathbf{y}$, $\tilde{\mathbf{v}}_\mathbf{z}$), in each axis. For RGB-D error modeling, we start testing with Machine and Deep Learning multivariate regression algorithms. We provide accuracy for the errors predictions using the best methods.

A. Modeling Errors

We apply the ML and DL algorithms on **RMSE-Model/hall** and **RPE-Model/hall** datasets, evaluating Linear Regression (LR), Lasso Regression (LASSO), Elastic Net (EN), K-Neighbors Regressor (KNNR), Decision Tree Regressor (DTR), Support Vector Regressor (SVR), Ada Boost Regressor (ABR), Gradient Boosting Regressor (GBMR), Random Forest Regressor (RFR), Extra Trees Regressor (ETR), and also the use of a Multi-Layer Perceptron (MLP). Using data exploration, we could identify several relevant features of **RMSE-Model/hall** and **RPE-Model/hall** such as: a) RMSE

increases with vibration increase; b) RMSE decreases with distance increase; c) RMSE decreases with velocity increase; d) RPE increases with velocity increase; e) RPE increases with vibration increase; f) and RPE decreases with distance increase.

We provide a thoroughly $cross\text{-}validation$ analysis for the regression methods with the $scikit\text{-}learn$ library. The data in **RMSE-Model/hall** and **RPE-Model/hall** are split into two parts, 75% for training and 25% for validation. After $cross\text{-}validation$, we found that the best ML algorithms to model the 3D reconstruction RMSE and the camera RPE are S-SVR and LR, respectively. The accuracy of S-SVR is $\pm$0.31 m, i.e., 36% of the 0.87 m that is the mean of all data. The LR accuracy is $\pm$8.2e-04 m/frame, i.e., 43% of the mean 0.0019 m/frame. This value means that the ML algorithms did not do a good prediction job. Thus, to improve the ML results, we experiment with the MLP neural network. First, we define the network structure through varying the number of neurons in the first hidden layer, adding new hidden layers and varying the number of neurons, and by varying the training $epochs$ and using data transformations. After these tests, the result is two MLP neural networks $\mathcal{N}_{RMSE_{X,Y,Z}}$ and $\mathcal{N}_{RPE_{x,y,z}}$ for the multivariate regression of the 3D reconstruction and the camera positioning errors, respectively.

B. Error Predictions Evaluation

The columns $\tilde{\mathbf{z}}$, $\tilde{\mathbf{s}}$ and $\tilde{\mathbf{v}}$ of **RMSE-Eval/hall** and **RPE-Eval/hall** are the new inputs of the neural network $\mathcal{N}_{RMSE_{\hat{\mathbf{f}}}}$ and $\mathcal{N}_{RPE_{\hat{\mathbf{f}}}}$, respectively. The predictions are reported in Table I and Table II. Analyzing Table I and II, it can be seen that the 3D reconstruction error is approximately $\pm$0.023 m, i.e. approximately $\pm$1% of the nominal value. The error in the prediction of new camera RPE values is approximately $\pm$2.66e-05 m/frame, i.e. approximately $\pm$2.5% of the nominal value.

C. Practical Applications

To show the usefulness of our method, we start presenting a point cloud correction processing. Specifically, only the spatial positioning (only translation) of the 3D points is corrected by adding the reconstruction errors predicted by the $\mathcal{N}_{RMSE_{\hat{\mathbf{f}}}}$ network. Figure 4a show the cloud before and after correction. It can be seen how each point position is relocated to a position nearby the control points (green points).

A second application on camera odometry (only translation) correction is devised, with a random camera trajectory with their respective velocity and vibration. Figure 4b shows the camera trajectory in each axis (light blue lines) and its correction (light green line) by adding the RPE obtained by $\mathcal{N}_{RPE_{\hat{\mathbf{f}}}}$ network predictions.

VI. Conclusions

Our work proposal includes a versatile framework for estimating error accuracy in 3D camera pose determination and 3D reconstruction from data provided by RGB-D sensors coupled to mobile robotic platforms. The approach is based on a new methodology for determining ground truth, from which the use of neural networks is utilized to approximate the predictions when the robot is operating in its environment. In general, the results in errors prediction confirm our hypothesis that it is possible to model the errors in RGB-D camera measurements as a function of other magnitudes such as distance, velocity, and vibration; and above all, it opens the door to new models based on other physical magnitudes.

The training of $\mathcal{N}_{RMSE_{\hat{\mathbf{f}}}}$ and $\mathcal{N}_{RPE_{\hat{\mathbf{f}}}}$, was carried out using data captured within a stage with a flat marble floor. The vibration level of the camera is related to the floor surface, i.e., the neural networks presented could give wrong predictions when there is a vibration magnitude outside the levels used for their training. Thus, we plan to investigate how our method behaves on the presence of larger vibrations (i.e. uneven surfaces).

References

[1] Y. Siddiqui, J. Thies, F. Ma, Q. Shan, M. Niesner, and A. Dai, "Retrievalfuse: Neural 3d scene reconstruction with a database," in *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*. Los Alamitos, CA, USA: IEEE Computer Society, oct 2021, pp. 12 548–12 557. [Online]. Available: https://doi.ieeecomputersociety.org/10.1109/ICCV48922.2021.01234

[2] Z. Murez, T. van As, J. Bartolozzi, A. Sinha, V. Badrinarayanan, and A. Rabinovich, "Atlas: End-to-end 3d scene reconstruction from posed images," in *Computer Vision – ECCV 2020*, A. Vedaldi, H. Bischof, T. Brox, and J.-M. Frahm, Eds. Cham: Springer International Publishing, 2020, pp. 414–431.

[3] A. M. Brito-Junior, A. D. Dória-Neto, J. D. Melo, and L. M. G. Gonçalves, "An adaptive learning approach for 3-d surface reconstruction from point clouds," *IEEE Transactions on Neural Networks*, vol. 19, no. 6, pp. 1130–1140, 2008.

[4] M. Oe, T. Sato, and N. Yokoya, "Estimating camera position and posture by using feature landmark database," in *Image Analysis*, H. Kalviainen, J. Parkkinen, and A. Kaarna, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2005, pp. 171–181.

[5] R. Juarez-Salazar, L. N. Gaxiola, and V. H. Diaz-Ramirez, "Single-shot camera position estimation by crossed grating imaging," *Optics Communications*, vol. 382, pp. 585–594, 2017. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0030401816307271

[6] A. A. S. Souza and L. M. G. Gonçalves, "2.5-dimensional grid mapping from stereo vision for robotic navigation," in *2012 Brazilian Robotics Symposium and Latin American Robotics Symposium*, 2012, pp. 39–44.

[7] L. Ortiz, V. Cabrera, and L. Gonçalves, "Depth Data Error Modeling of the ZED 3D Vision Sensor from Stereolabs," *ELCVIA Electronic Letters on Computer Vision and Image Analysis*, vol. 17, no. 1, pp. 1–15, 2018.

[8] L. Rocha and L. Gonçalves, "An overview of three-dimensional videos: 3d content creation, 3d representation and visualization," in *Current Advancements in Stereo Vision*, A. Bhatti, Ed. Rijeka: IntechOpen, 2012, ch. 6, pp. 1–12. [Online]. Available: https://doi.org/10.5772/50177

[9] B. Silva, A. Burlamaqui, and L. Gonçalves, "On monocular visual odometry for indoor ground vehicles," in *2012 Brazilian Robotics Symposium and Latin American Robotics Symposium*, 2012, pp. 220–225.

[10] E. V. Cabrera, L. E. Ortiz, B. M. F. d. Silva, E. W. G. Clua, and L. M. G. Gonçalves, "A Versatile Method for Depth Data Error Estimation in RGB-D Sensors," *Sensors*, vol. 18, no. 9, 2018.

[11] S. Okazaki, T. Tanaka, S. Kaneko, H. Takauji, N. Kochi, and M. Yamada, "Modeling stereo measurement error by considering camera vibration," in *2010 Int. Symp. on Optomechatronic Technologies*, Oct 2010, pp. 1–6.

[12] A. Lavatelli and E. Zappa, "Modeling uncertainty for a vision system applied to vibration measurements," *IEEE Transactions on Instrumentation and Measurement*, vol. 65, no. 8, pp. 1818–1826, 2016.

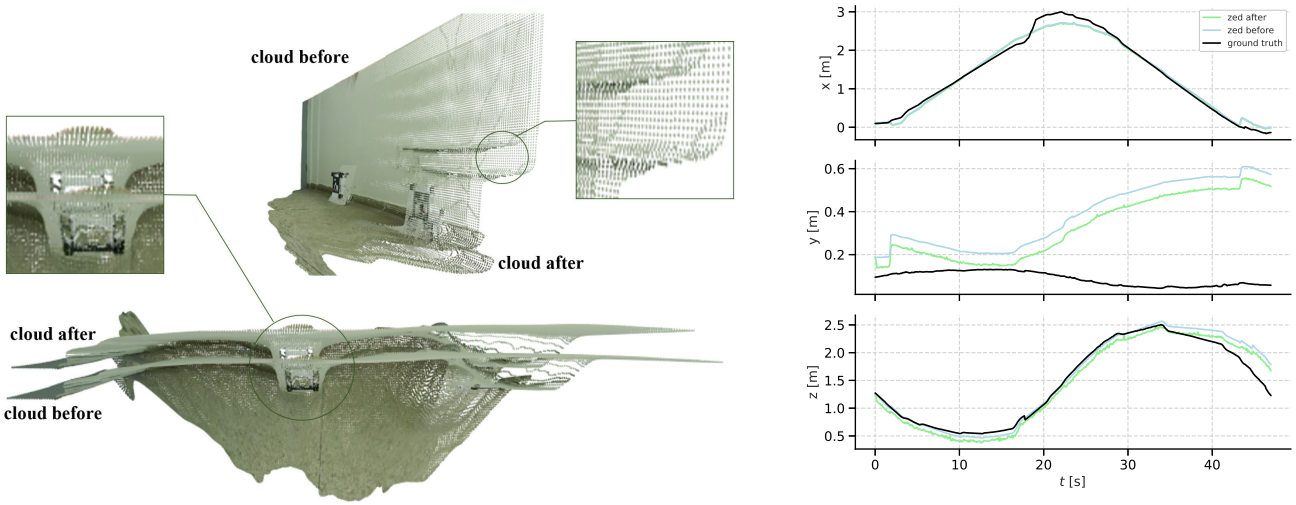
[13] P. J. Besl and N. D. McKay, "A method for registration of 3-D shapes," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 14, no. 2, pp. 239–256, Feb 1992.

TABLE I: $\mathcal{N}_{RMSE_{\hat{\mathrm{f}}}}$ predictions for some samples of the dataset **RMSE-Eval/hall**.

Distance [m]			**Velocity [m/s]**			**Vibration [g]**			$RMSE_{\hat{\mathrm{f}}}$ **Nominal [m]**			$RMSE_{\hat{\mathrm{f}}}$ **Predicted [m]**			**Prediction Error [m]**		
x	y	z	x	y	z	x	y	z	x	y	z	x	y	z	x	y	z
-0.45	0.08	11.29	0.4	0.4	0.1	-0.02	-1.03	-1.95e-02	0.11	0.17	0.79	0.11	0.17	0.80	2.58e-04	1.07e-03	4.07e-03
-0.44	0.08	11.37	0.1	0.6	0.1	-0.03	-1.04	-7.80e-03	0.14	0.21	0.84	0.14	0.21	0.84	5.02e-05	1.92e-04	5.61e-04
-0.44	0.13	11.54	0.2	0.4	0.2	-0.04	-1.03	-1.17e-02	0.13	0.17	0.59	0.13	0.17	0.59	8.71e-04	8.90e-04	6.88e-03
-0.42	0.14	11.88	0.6	0.2	0.3	-0.04	-1.02	-1.17e-02	0.08	0.22	0.66	0.08	0.22	0.66	3.86e-04	1.41e-03	6.62e-03
-0.33	0.15	12.91	0.4	0.1	0.1	-0.02	-1.03	-1.17e-02	0.11	0.22	0.74	0.11	0.22	0.75	2.03e-04	1.10e-03	5.94e-03
-0.28	0.16	12.92	0.1	0.1	0.1	-0.03	-1.03	-7.80e-03	0.15	0.22	0.74	0.15	0.22	0.74	2.68e-04	1.39e-03	1.78e-03

TABLE II: $\mathcal{N}_{RPE_{\hat{\mathrm{f}}}}$ predictions for some samples of the dataset **RPE-Eval/hall**.

Distance [m]			**Velocity [m/s]**			**Vibration [g]**			$RPE_{\hat{\mathrm{f}}}$ **Nominal [m/frame]**			$RPE_{\hat{\mathrm{f}}}$ **Predicted [m/frame]**			**Prediction Error [m/frame]**		
x	y	z	x	y	z	x	y	z	x	y	z	x	y	z	x	y	z
-0.95	0.16	14.13	0.3	0.2	0.2	-0.04	-1.02	3.90e-03	4.59e-04	4.05e-04	1.99e-03	3.74e-04	3.62e-04	1.85e-03	8.47e-05	4.33e-05	1.40e-04
-0.95	0.16	14.26	0.4	0.2	0.1	-0.03	-1.03	-1.56e-02	5.27e-04	4.01e-04	1.85e-03	4.42e-04	4.23e-04	1.70e-03	8.51e-05	2.22e-05	1.49e-04
-0.72	0.16	14.40	0.3	0.2	0.2	-0.03	-1.02	-1.95e-02	4.75e-04	4.11e-04	1.96e-03	3.90e-04	3.90e-04	1.75e-03	8.55e-05	2.05e-05	2.13e-04
-0.43	0.17	14.43	0.1	0.2	0.4	-0.03	-1.04	-3.90e-03	4.45e-04	5.28e-04	2.08e-03	3.72e-04	5.12e-04	2.24e-03	7.25e-05	1.62e-05	1.60e-04
-0.31	0.18	14.73	0.1	0.2	0.2	-0.03	-1.04	-1.56e-02	4.53e-04	4.82e-04	1.65e-03	4.14e-04	5.10e-04	1.81e-03	3.86e-05	2.82e-05	1.60e-04
-0.26	0.19	14.74	0.1	0.2	0.3	-0.04	-1.03	-2.73e-02	5.49e-04	5.64e-04	2.38e-03	4.62e-04	5.26e-04	2.17e-03	8.64e-05	3.75e-05	2.15e-04



(a) (b)

Fig. 4: (**a**) Point cloud correction. (**b**) Camera odometry (translation) correction.

[14] D. Holz, A. E. Ichim, F. Tombari, R. B. Rusu, and S. Behnke, "Registration with the Point Cloud Library: A Modular Framework for Aligning in 3-D," *IEEE Robotics Automation Magazine*, vol. 22, no. 4, pp. 110–124, 2015.

[15] K. S. Arun, T. S. Huang, and S. D. Blostein, "Least-Squares Fitting of Two 3-D Point Sets," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. PAMI-9, no. 5, pp. 698–700, 1987.

[16] B. K. P. Horn, "Closed-form solution of absolute orientation using unit quaternions," *Journal of the Optical Society of America A*, vol. 4, no. 4, pp. 629–642, 1987.

[17] M. W. Walker, L. Shao, and R. A. Volz, "Estimating 3-D location parameters using dual Number quaternions," *CVGIP: Image Understanding*, vol. 54, pp. 358–367, 1991.

[18] L. E. Ortiz-Fernandez, E. V. Cabrera-Avila, B. M. F. d. Silva, and L. M. G. Gonçalves, "Smart artificial markers for accurate visual mapping and localization," *Sensors*, vol. 21, no. 2, 2021. [Online]. Available: https://www.mdpi.com/1424-8220/21/2/625

[19] R. Kümmerle, G. Grisetti, H. Strasdat, K. Konolige, and W. Burgard, "G2o: A general framework for graph optimization," in *2011 IEEE International Conference on Robotics and Automation*, May 2011, pp. 3607–3613.

[20] S. Yang, B. Bhanu, and A. I. Mourikis, "Error model for scene reconstruction from motion and stereo," in *IEEE Comp. Soc. Conf. on Computer Vision and Pattern Recognition (Workshops)*, 2010, pp. 70–77.

[21] T. Zhang and T. Boult, "Realistic stereo error models and finite optimal stereo baselines," in *2011 IEEE Workshop on Applications of Computer Vision (WACV)*, Jan 2011, pp. 426–433.

[22] S. Z. Bo Jin, Lijun Zhao, "Error modelling of depth estimation based on simplified stereo vision for mobile robots," *Computer Modelling and New Technologies*, vol. 18, no. 4, pp. 450–454, 2014.

[23] P. O. Mikko Kytö, Mikko Nuutinen, "Method for measuring stereo camera depth accuracy based on stereoscopic vision," *Proc. SPIE*, vol. 7864, pp. 7864–7864–9, 2011.

[24] Y. Xu, Y. Zhao, F. Wu, and K. Yang, "Error analysis of calibration parameters estimation for binocular stereo vision system," in *2013 IEEE Int. Conf. on Imaging Systems and Techniques (IST)*, 2013, pp. 317–320.

[25] W. Sankowski, M. Włodarczyk, D. Kacperski, and K. Grabowski, "Estimation of measurement uncertainty in stereo vision system," *Image and Vision Computing*, vol. 61, pp. 70 – 81, 2017.

[26] N. Kehtarnavaz and W. Sohn, "Error Analysis of Camera Movements in Stereo Vehicle Tracking Systems," *Computer Vision and Image Understanding*, vol. 62, no. 3, pp. 347 – 359, 1995.

[27] G. Kocak, S. Yamamoto, and T. Hashimoto, "Analyzing Influence of Ship Movements on Stereo Camera System Set-up on Board Ship," *Marine Engineering*, vol. 47, no. 6, pp. 888–895, 2012.

[28] S. Okazaki, T. Tanaka, S. Kaneko, H. Takauji, N. Kochi, and M. Yamada, "Reliability Analysis of Object Position Measured by Motion Stereo Considering Camera Vibration," *IEEJ Transactions on Industry Applications*, vol. 132, no. 10, pp. 942–950, 2012.

[29] Y. Wang, J. Zhang, and H. Deng, "Error analysis of dynamical measurement system based on binocular vision," *APCOM*, 2013.