2D-3D Camera Fusion for Visual Odometry in Outdoor Environments

Danda Pani Paudel, Cédric Demonceaux, Adlane Habed, Pascal Vasseur, Inso Kweon

2D-3D Camera Fusion for Visual Odometry in Outdoor Environments

Danda Pani Paudel[1] Cédric Demonceaux[1] Adlane Habed[2] Pascal Vasseur[3] and In So Kweon[4]

Abstract— **Accurate estimation of camera motion is very important for many robotics applications involving SfM and visual SLAM. Such accuracy is attempted by refining the estimated motion through nonlinear optimization. As many modern robots are equipped with both 2D and 3D cameras, it is both highly desirable and challenging to exploit data acquired from both modalities to achieve a better localization. Existing refinement methods, such as Bundle adjustment and loop closing, may be employed only when precise 2D-to-3D correspondences across frames are available. In this paper, we propose a framework for robot localization that benefits from both 2D and 3D information without requiring such accurate correspondences to be established. This is carried out through a 2D-3D based initial motion estimation followed by a constrained nonlinear optimization for motion refinement. The initial motion estimation finds the best possible 2D-to-3D correspondences and localizes the cameras with respect the 3D scene. The refinement step minimizes the projection errors of 3D points while preserving the existing relationships between images. The problems of occlusion and that of missing scene parts are handled by comparing the image-based reconstruction and 3D sensor measurements. The effect of data inaccuracies is minimized using an M-estimator based technique. Our experiments have demonstrated that the proposed framework allows to obtain a good initial motion estimate and a significant improvement through refinement.**

I. INTRODUCTION

In this paper we aim to fuse calibrated synchronized 2D and 3D moving cameras for a better estimation of their motion. Accurate motion estimation is of prime importance in visual Simultaneously Localization and Mapping (vSLAM). An accurate environment map is generally required for an accurate localization. In turn, building an accurate environment map is not possible without an accurate localization, hence, making it a paradoxical "chicken and egg" problem.

With the relatively recent proliferation of affordable 3D and 2D sensors, many mobile robots are, or can easily be, equipped with both 2D and 3D cameras [1][2][4][5][6]. Most of such robots localize themselves using the Iterative Closest Point (ICP) algorithm (or one of its variants). Some robots also use 2D images for localization whereas the mapping is done using the 3D sensors. Long run outdoor localization based on 3D information alone is difficult mainly because of the unreliable 3D feature descriptors and local minima traps (typical to ICP). In the 2D-2D case, however, the development of reliable 2D image feature descriptors (such as SIFT), 2D-to-2D matching has become more trustworthy. Unfortunately, 2D features alone may not allow to compute the motion up to desired accuracy. Furthermore, finding precise 2D-to-3D correspondences based on 2D features matching is not trivial even for a calibrated setup. Localization based on such 2D-to-3D correspondences and 2D-2D based refinement may suffer from significant error accumulation. One example of such error accumulation is shown in Fig. 1. This error is usually minimized by loop closing techniques as described in [3]. However, in particular when robots travel long distances, loop closing, if ever possible, may not adequately compensate for error accumulation thus leaving visible artifacts in the map. This demands robots to make small and frequent loops so that the accumulated error remains under control. In practice, making such small loops while building large maps is undoubtedly a burden for the task at hand and often impossible. While incorporating information from extra sensors such as GPS has been proposed [6][8], it is often argued that such information is neither accurate nor reliable enough. It is highly recommended, when building large maps, to perform the loop closing with large real loops, whenever possible, thus reducing the accumulation error. Consequently, the robots moving around large structures require a very accurate localization: good localization makes the paradoxical vSLAM problem less difficult.

Visual odometry is generally carried out by relying on 2D-2D, 3D-3D, or 2D-3D information. 2D-2D based methods typically track features in monocular or stereo images and estimate the motion between them [9][10]. Some of these methods improve the localization accuracy by simultaneously processing multiple frames, while using Bundle Adjustment (BA) for refinement. Some other methods obtain the motion parameters by registering images such that the photometric error between them is minimized [11], [12]. For the same purpose, most 3D-3D based methods use ICP or its variants [13][20][14] between conjugately acquired point clouds obtained from the 3D camera [18][17]. However, ICP-based methods are computationally expensive due to the calculation of the nearest neighbors for every point at each iteration. Both of these methods use the information from either camera only and, hence, do not fully exploit all the available information. Recent works [19][15] propose the use of information provided from both cameras during the process of localization. The work in [19] refines the camera pose obtained from Structure-from-Motion (SfM) using an extra constraint of a plane-induced homography via scene planes. This method provides a very good insight for a possibility to improve the camera pose when the partial 3D is known. However, it uses only the information from planes that are

This research has been funded by an International Project NRF-ANR DrAACaR: ANR-11-ISO3-0003.

[1]Le2i UMR 6306, CNRS, University of Burgundy, France
[2]ICube UMR 7357, CNRS, University of Strasbourg, France
[3]LITIS EA 4108, University of Rouen, France
[4]Robotics and Computer Vision Lab, KAIST, Korea

in the scene. The methods presented in [17][15][16] have been tested in indoor environments mainly with a Kinect sensor. Extension of these methods to outdoor environments with possibly different kinds of 3D cameras is not trivial due to various unhandled situations that may arise. Typical issues arising in outdoor scenes and/or different camera setups occur, for example, when 2D and 3D cameras do not share the exact same field of view, when the 3D points are sparse (as opposed to pixel-to-pixel mapping of RGB-D cameras), in the absence of required scene structures, and in the event of low frame rates and/or large displacements of the cameras. Note that other 2D-3D based existing refinement methods, such as Bundle adjustment and loop closing, are not applicable under these circumstances because they require precise 2D-to-3D correspondences across frames.

In this work, we first propose a complete framework for visual odometry of 2D-3D camera system in an outdoor environment addressing the above mentioned difficulties. This framework computes the pose by localizing a set of cameras at once with respect to the 3D scene acquired in the previous frame using a minimum of three corresponding points among all the views. We also propose a constrained nonlinear optimization framework that further refines this pose. The first step of our method uses only the known part of the scene whereas our refinement process uses the constraints that arise from the unknown part of the scene as well. Unlike [19], our method makes no prior assumption regarding the geometry of the scanned scene. Furthermore, the proposed method differs from [23] as it has been specifically designed for synchronized 2D-3D cameras in outdoor setup.

Our paper is organized as follows: we introduce the background and the used notations in Section II. We formulate the optimization problem to obtain the optimal odometry parameters in Section III. The solution to this problem is presented in the form of an algorithm in the same section. In Section IV, experiments with two real datasets are presented and discussed. Section V concludes our work.

II. Notation and Background

The setup consists of a 3D scanner and multiple calibrated cameras as shown in Fig. 2. At any given instant, the 3D scanner scans the scene points $X_k^1, k = 1 \dots p$ in its coordinate frame O^1. At the same time, calibrated cameras with known extrinsic parameters $R_i|t_i, i = 1 \dots m$ capture m images, from which a set of 2D feature points are extracted. Let $x_{ij}^1, j = 1 \dots n$ represent those feature points in the i^{th} image. $P(R, t, X)$ is the projection function that maps a point X to its 2D counterpart in the images captured from $R|t$. When the system moves by $R'|t'$ to next position, the corresponding variables are represented by similar notations with change in superscript. If x_{ij}^1 and $x_{ij}^2, j = 1 \dots n$ are the corresponding sets of feature points in two consecutive images taken by i^{th} camera, their 2D-to-3D correspondences are specified by a function ϕ. Let $\phi_i(j)$ be a function that maps each pair of 2D points $x_{ij}^1 \leftrightarrow x_{ij}^2$, to the corresponding 3D point X_k^1. Every rotation matrix R is represented by a 4×1 vector of quaternions q unless

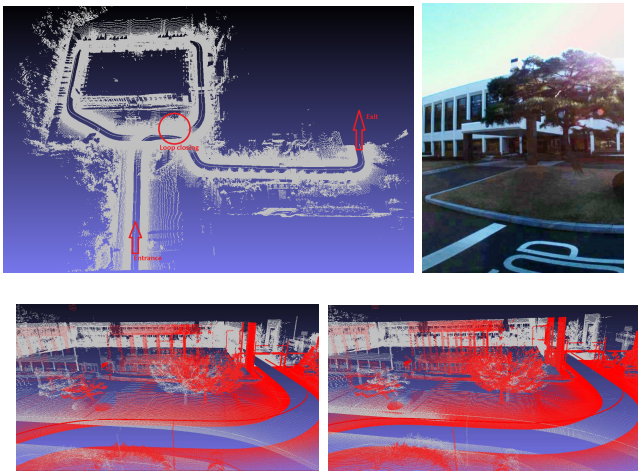

Fig. 1: An example of error accumulation around a loop: Map built by a Laser-Camera system around a large structure (top-left). Image taken at a loop closing point with only one tree at the corner (top-right). Map built before (red) and after (white) the visit around the loop using 2D-2D based refinement [6] (bottom-left). Refined map obtained using our method (bottom-right). The scans of the same tree distant before come significantly closer after refinement.

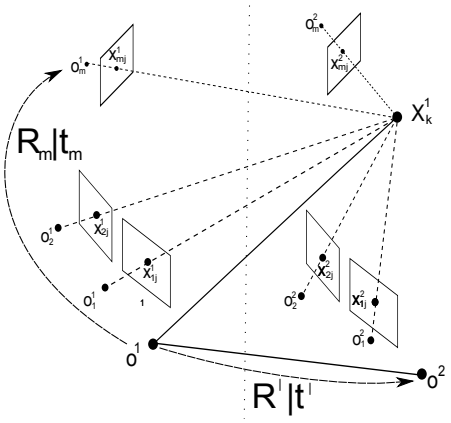

Fig. 2: Ray diagram of the experimental setup.

mentioned otherwise. Similarly, q' for R'. Both 3D and 2D points are represented by 3×1 vectors, the latter being the homogeneous representation in camera coordinate system.

III. 2D-3D Odometry

In this section, we establish the relationships between pairs of image points in two views and the scene points acquired from the first one. Using these relationships, we propose an optimization framework whose optimal solution is the required odometry parameters. A complete algorithm for solving this optimization problem is also discussed. The proposed method deals with the case in which the 2D cameras and 3D sensors are synchronized and 2D-to-2D correspondences between the pairs of images acquired by the same camera are known.

A. 2D-3D based Localization

It is trivial to find the 2D-to-3D correspondences, $X_k^1 \leftrightarrow P(R_m, t_m, X_k^1)$ in one frame. However, we need cross-frame correspondences to estimate the motion $R'|t'$. Such correspondences can be obtained by matching the 2D feature points between images. Note that, most $P(R_m, t_m, X_k^1)$, when considered as feature points, are unlikely to result in reliable feature descriptors for matching. Therefore, we extract a separate set of 2D feature points to obtain better 2D-2D correspondences $x_{ij}^1 \leftrightarrow x_{ij}^2$. Motion estimation from these correspondences requires at least 5 points to compute the motion with an unknown scale. On the other hand, if we can find 2D-3D correspondences $x_{ij}^2 \leftrightarrow X_k^1$, it would require only 3 points to estimate the motion including the scale. In order to benefit from this, the required 2D-to-3D correspondences are computed for each image which is established by the mapping function $\phi_i(j)$ computed as

$$\phi_i(j) = \underset{k \in \{1,\dots,p\}}{\operatorname{argmin}} \; ||x_{ij}^1 - P(R_i, t_i, X_k^1)||, j = 1 \dots n. \tag{1}$$

It is important to notice that the correspondences obtained in this manner are not perfect. We make a strong consideration of this restriction while refining the estimated motion. The search required to minimize (1) can be performed using a KD-tree like structure where the projections of all 3D points build one tree in each image. The detected feature points traverse these trees in search for the best possible match. Once the required correspondences are obtained, the set of cameras in second frame can be localized with respect to the previously acquired 3D scene using the method presented in [7]. The advantage of using this method is that it requires a minimum of 3 correspondences among all the views and does not require a complex scene as demanded by ICP or SfM. For example, even a planar scene with sufficient texture can be processed. For low frame rates and/or large displacements, feature matching methods still work better than tracking them. Since only 3 correspondences are needed, finding them from already matched 2D-2D to sparse 3D is very much achievable in practice.

B. 2D-2D-to-3D based motion refinement

In the refinement process, we wish to minimize the sum of projection errors over all the computed 2D-to-3D correspondences. Hence, the cost to be minimized is defined as

$$\zeta(R', t') = \sum_{i=1}^{m} \sum_{j=1}^{n} ||x_{ij}^2 - P(R_i R', R' t_i + t', X_{\phi_i(j)}^1)||^2. \tag{2}$$

At the same time, the Essential matrix between two views of the same camera in different frames is expressed as

$$E_i(R', t') = [t_i']_\times R_i', \tag{3}$$

where $R_i'|t_i'$ is the pose of i^{th} camera in second frame with respect to the first one. It is related to $R'|t'$ as follows

$$\begin{pmatrix} R_i' & t_i' \\ 0 & 1 \end{pmatrix} = \begin{pmatrix} R_i & t_i \\ 0 & 1 \end{pmatrix} \begin{pmatrix} R' & t' \\ 0 & 1 \end{pmatrix} \begin{pmatrix} R_i & t_i \\ 0 & 1 \end{pmatrix}^{-1}. \tag{4}$$

Hence, the epipolar constraint that relates the points in two views of different frames can be written as

$$(x_{ij}^2)^T E_i(R', t') x_{ij}^1 = 0. \tag{5}$$

Note that the minimization of cost defined in (2) locates the set of cameras of the second frame with respect the 3D points cloud acquired in the first frame. Similarly, (5) localizes the second camera with respect to the first one. Theoretically, it can be seen that (2) and (5) are redundant. However, in the presence of noisy data and unknown correspondences, satisfying only the non-redundant condition does not necessarily satisfy the other. Recall that, the 2D-to-3D correspondences computed from (1) are not precise enough to obtain an accurate localization. Hence, localizing the cameras based only on the minimization of a cost function derived for such inaccurate data may simply destroy the relative pose. Satisfying only (5) does not make use of the known scene. Therefore, we choose to incorporate both equations in an optimization framework to obtain a better solution. For the optimization, the cost of (2) is minimized while imposing (5) as a constraint. It is important to notice that (5) makes use of the unknown part of the scene as well.

Basically, our problem is to localize a set of 2D cameras for known 2D-to-2D $(x_{ij}^1 \leftrightarrow x_{ij}^2)$ and unknown 2D-2D-to-3D $(x_{ij}^1 \leftrightarrow x_{ij}^2 \leftrightarrow X_{\phi_i(j)}^1)$ correspondences in a noisy environment. Hence, finding the optimal ϕ_i itself is part of the optimization process. The motion estimation based on computed ϕ_i is called 2D-3D registration. Finding better values of R' and t' from 2D-2D-to-3D relationship is called the camera pose refinement. Both registration and refinement processes in a common optimization framework is written as

$$\begin{aligned} \underset{q', t', \phi}{\min} \; & \sum_{i=1}^{m} \sum_{j=1}^{n} ||x_{ij}^2 - P(R_i R', R' t_i + t', X_{\phi_i(j)}^1)||^2, \\ \text{subject to} \; & (x_{ij}^2)^T E_i(R', t') x_{ij}^1 = 0, \\ & ||q'||^2 = 1, \; i = 1 \dots m, j = 1 \dots n. \end{aligned} \tag{6}$$

The optimization problem (6) considers that every image point has its corresponding 3D point in the scene. In practice, there could be two problems: (a) multiple 3D points lying on the back-projected ray from the second camera center through an image point. All such points satisfy the epipolar constraint and hence, lead to correspondence ambiguity, and (b) extra 2D or missing 3D points resulting invalid 2D-to-3D correspondences. We address both of these problems by assigning the weights derived from the scale histogram for each correspondence.

If $\widetilde{X}_{ij}$ is the two-view reconstruction obtained using the motion estimated from 2D-to-3D correspondences, the relative scale of reconstruction for known 3D-to-3D correspondences $\widetilde{X}_{ij} \leftrightarrow X_{\phi_i(j)}^1$ is computed as

$$s_i(j) = \frac{||R_i^T \widetilde{X}_{ij} - R_i^T t_i||}{||X_{\phi_i(j)}^1||}, \quad j = 1 \dots m. \tag{7}$$

Since the motion is estimated with true scales, in the ideal case $s_i(j) = 1 \; \forall i \in 1 \dots m, j \in 1 \dots n$. In practice, when

a combined histogram $H(u), u = 1 \dots b$ of these scales is built, it holds the highest number of samples in the bin corresponding to true scale (usually, close to 1). If that bin is $u_{\max}$, then the weights are distributed as follows:

$$w_i(j) = \begin{cases} 1 & s_i(j) \in H(u_{\max}) \\ 0 & \text{otherwise.} \end{cases} \tag{8}$$

Furthermore, the effect of data inaccuracies is reduced by introducing a robust estimation technique. Hence, the optimization problem (6) with robust estimation and histogram based weighting can be re-written as

$$\begin{aligned} \min_{q', t', \phi} \quad & \sum_{i=1}^{m} \sum_{j=1}^{n} w_i(j) \rho(||x_{ij}^2 - P(R_i R', R' t_i + t', X_{\phi_i(j)}^1)||), \\ \text{subject to} \quad & \rho((x_{ij}^2)^T E_i(R', t') x_{ij}^1) = 0, \\ & ||q'||^2 = 1, \ i = 1 \dots m, j = 1 \dots n. \end{aligned} \tag{9}$$

where $\rho(x)$ is Tukey bi-weighted potential function. For a threshold of ξ, it is defined as

$$\rho(y) = \begin{cases} \frac{y^6}{6} - \frac{\xi^2 y^4}{2} + \frac{\xi^4 y^2}{2} & for\ |y| < \xi \\ \frac{\xi^6}{6} & \text{otherwise} \end{cases} \tag{10}$$

whose influence function is $\psi(y) = y\left(\xi^2 - y^2\right)^2$ for $|y| < \xi$ and 0 otherwise.

Note that, the cost is derived only for the known part of the scene. However, the constraint includes the unknown part of the scene as well. The optimal odometry parameters are obtained by iteratively solving this optimization problem. First, 2D-to-3D correspondences are found using (1) and the method [7] is used to estimate intial R' and t'. Then, the rest is a constrained nonlinear optimization problem whose local optimal solution can be obtained by iteratively re-weighted least-squares (IRLS) technique. Each iteration of IRLS uses interior-point method to solve the constrained nonlinear least-squares problem. In our implementation, we have relaxed the strict equality of constraints to avoid the infeasibility that would arise due to noisy data.

C. The algorithm

For known extrinsic parameters $R_i|t_i, i = 1 \dots m$ and calibrated cameras, the proposed method works in two steps. Both steps are carried out for each pair of consecutive frames l and $l+1$ as described in Algorithm 1.

D. Normalization and pose recovery

To avoid any numerical issues arising due to the disproportionate measurement of different systems, the 3D points are scaled and transformed such that their centroid remains within one unit away from the 3D camera coordinate system. All other translation terms are also normalized accordingly, i.e. if $\lambda = (\sum_{l=1}^{p} ||X_l^1||)/p$, they are scaled to $\lambda t_i, i = 1 \dots m$. Note that, the knowledge of λ is sufficient to recover the estimated motion with true scale. We also normalize the data during robust estimation i.e. y in (10) is scaled with twice of its median value and ξ is set to 1, whenever used.

Algorithm 1 2D-3D Odometry

Extract and match feature points to obtain $x_{ij}^l \leftrightarrow x_{ij}^{l+1}$. Iterate over following two steps until convergence.

1) **2D-3D registration**:
   For each Camera $i = 1 \dots m$,
   a) Compute $P(R_i, t_i, X_k^l), k = 1 \dots p$ and build a KD-tree.
   b) Find 2D-to-3D correspondences maps $\phi_i(j), j = 1 \dots n$ using (1).

   Using all Cameras: Perform 2D-3D based RANSAC and estimate $R_0'|t_0'$ using [7].
2) **2D-2D-to-3D based refinement**: Starting from $R_0'|t_0'$, iterate until convergence,
   a) Reconstruct the scene $\widetilde{X}_{ij}^l, j = 1 \dots n$ and compute scales $s_i(j)$ for each point.
   b) Build a combined scale histogram $H(u), u = 1 \dots b$ for all cameras.
   c) Compute weights $w_i(j), j = 1 \dots n$ using $H(u)$.
   d) Update the pose by optimizing (9) for known $\phi_i(j)$ obtained from step 1(c).

IV. Experiments

We have tested our method using two different real datasets. Both datasets were acquired by a moving vehicle equipped with a laser-camera system. However, these two setups greatly differ from one another. We have used SURF descriptor based matching to obtain the 2D-to-2D correspondences. The constrained nonlinear least-squares optimization problem is solved by using MATLAB-R2012a Optimization Toolbox with interior-point method.

KAIST Dataset: We conducted our first experiments using data obtained from a Laser-Camera system dedicated to reconstructing very large outdoor structures. This system uses two 2D laser scanners and four 2D cameras which are synchronized and calibrated for both intrinsic and extrinsic parameters. Laser scanners used here provide a wide angle of view of the scanning plane so that the system can observe tall objects as well as the ground making its suitable to scan the environment from a close distance. The 3D map (reconstruction) of the environment is made by collecting these 2D scans at the proper location. Therefore, this system requires a very precise localization for a good reconstruction. Extrinsic parameters of 2D cameras were estimated by laser points and a pattern-based calibration method. However, it still possesses the mean projection error of about 0.5 pixels. The interested reader may refer to [6] for details regarding the experimental setup. The dataset we have tested is a continuous trip of the Laser-Camera scanning system within the compound of KAIST (Korea) for a distance of about 3 KM. The system made seven different loops during its travel. The original reconstruction and the loops are shown in Fig. 3. The lengths of the loops, as shown in Table I, range from about 200 meters to 1.5 KM. Each camera captured 480×640 pix. images with a rate of about 20 frames/sec.

Loop	Size (m)	Bok *et al.* (m)	Our method (m)
1	351.76	4.063	1.548
2	386.38	4.538	1.469
3	224.37	4.765	4.398
4	242.87	1.696	1.077
5	931.14	3.884	2.858
6	1496.4	7.182	6.381
7	546.05	5.502	2.115

TABLE I: Loop size and loop closing errors in meters for Bok *et al.* [6] and our method.

The 2D-to-2D correspondences are computed between images escaping each 10 frames. The original reconstruction obtained by Laser-Camera system was used as the required 3D information for our method. Note that this reconstruction was not very accurate. Nevertheless, we were still able to refine the motion using such inaccurate data.

The qualitative and quantitative results are presented in Fig. 4 and Table I respectively. The errors were computed by performing the ICP between two points clouds captured at the loop closing point before and after the loop travel. Note that, loop closing methods are not applied to the presented results. Our goal is to obtain a better localization so that it would be suitable for the loop closing methods. We strongly believe that the localization with such accuracy can be a very suitable input for loop closing. Our experiments clearly show significant improvements in loop closing errors by our method for all the loops tested. Since, most of the loop closing methods used in practice provide only the local optimal solution; these improvements contribute to their convergence to the desired one. It can also be seen that the error reduction is independent of the loop length. In fact, the improvement is dependent upon the quality of feature points.The remaining residual error is the combined effect of the errors in calibration, matching, and measurements.

To analyze reconstruction accuracy, we fitted the surface on the reconstructed points cloud using an algorithm that we have developed in-house. This algorithm takes advantage of the camera motion and the order of scanned points. The reconstructed surface was mapped with texture from the same images that were used for localization. The textured scene with its various stages is shown in Fig. 5 for only one side of the reconstruction around the first loop (about 350 meters). This part of the reconstruction consists of about 1.3×10^6 3D points and 2.5×10^6 triangles.

KITTI Dataset: The proposed method was also tested on the benchmark dataset available at (http://www.cvlibs.net/datasets/kitti/). The details of the experimental setup is described in [22]. We have used the stereo pair of gray images and the 3D data scanned from a Velodyne laser scanner. The results obtained before and after refinement for 5 different sequences were compared against the provided ground truth. Errors in rotation and translation were computed by using the evaluation code provided along with the dataset which uses the ground truth obtained using GPS and other odometry sensors. Although this ground truth might not be very accurate for local poses

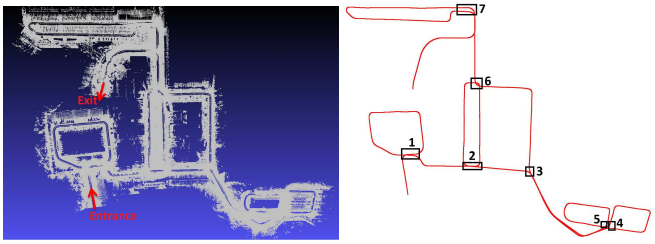

Fig. 3: Large map reconstructed using Laser-Camera system in a single trip shown with starting and end points (left). Closed loops made during the travel. Boxes shown are the loop closing locations of seven different loops (right).

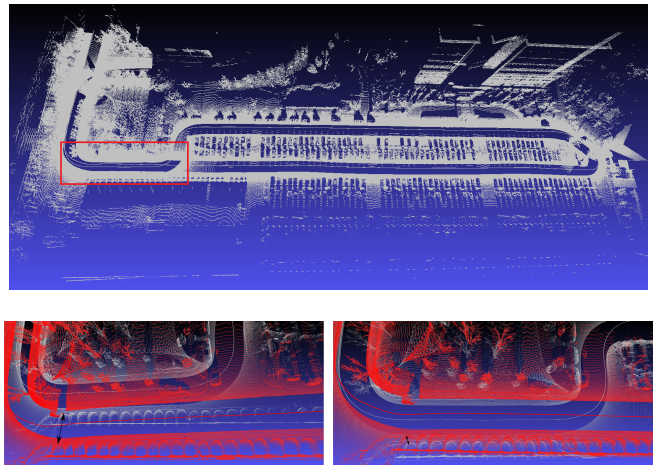

Fig. 4: Results similar to Fig. 1 for seventh Loop. Reconstruction with a red box at the loop closing location (top), obtained using Bok *et al.* (bottom-left) and our method after refinement (bottom-right). The double sided arrows show the gap between two different reconstructions of the same scene.

comparison, it is relevant over a long sequence due to no error accumulation process. Therefore, the errors were measured at the sequence steps of (100,200,...,800) and are presented in Table II. Fig. 6 shows the map obtained for the fifth sequence. A close observation shows that the localization before the refinement is already quite satisfactory. Its further refinement makes the result very close to the ground truth itself. Here again, the results are presented without the loop closing.

V. Conclusion

A method to fuse the information from 2D and 3D cameras for outdoor visual odometry has been proposed. Our demonstration with two different datasets show the possibility of estimating accurate motion of 2D-3D camera system even when the 3D scene is acquired up to some inaccuracies. Minimization of 3D projection errors while enforcing the relationship between images is key for such accuracy. An extension of our method to multi-frame processing is likely to improve the results further.

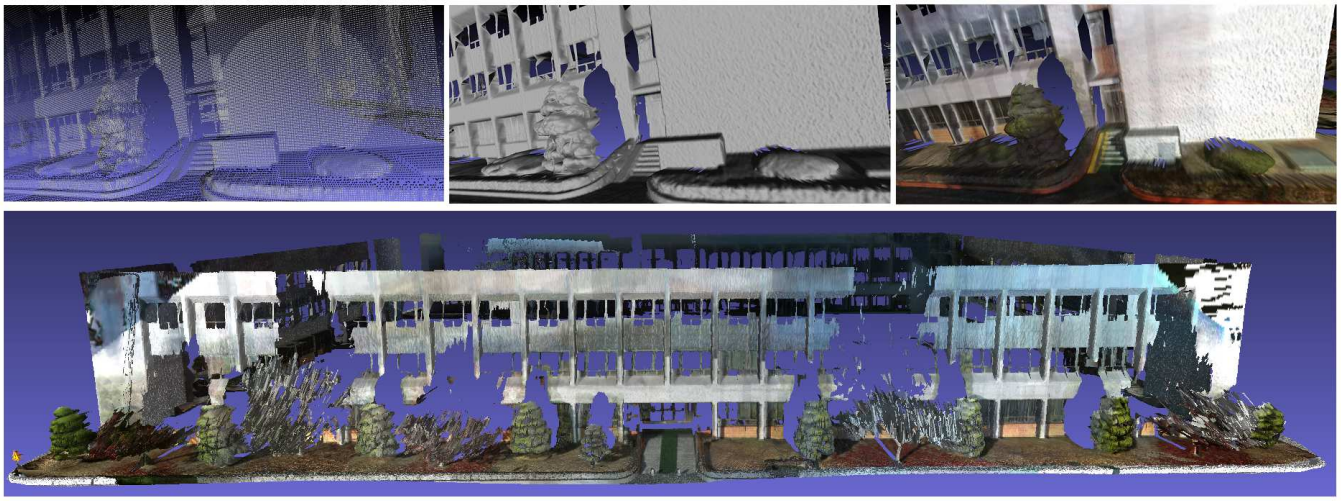

Fig. 5: Surface reconstruction and texture mapping showing the accuracy of localization. Reconstructed 3D, fitted surface, and texture mapping in a close view (top row, left to right). Texture mapping of the structure scanned around loop 1 (bottom).

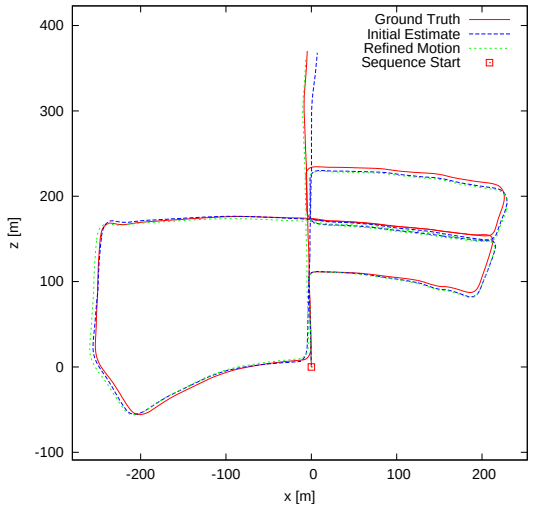

Fig. 6: Map built by our method (Initial Estimate and Refined Motion) vs. Ground Truth for the fifth sequence.

Sq.N	N.Frames	Initial Estimate		Refined	
		$\Delta T(\%)$	$\Delta R(^\circ/m)$	$\Delta T(\%)$	$\Delta R(^\circ/m)$
3	801	1.6774	0.000432	1.6398	0.000216
5	2761	1.9147	0.000245	1.8679	0.000162
7	1101	2.3410	0.000231	1.5689	0.000192
8	4071	2.3122	0.000447	1.9799	0.000196
9	1591	1.7562	0.000270	1.5604	0.000197

TABLE II: Translation (ΔT) and Rotation (ΔR) errors in Initial and Refined results for five different sequences.

REFERENCES

[1] D. Holz, C. Lorken, and H. Surmann, Continuous 3D sensing for navigation and SLAM in cluttered and dynamic environments, ICIF, 2008.

[2] J.W. Weingarten, G. Gruener, and R. Siegwart, A state-of-the-art 3D sensor for robot navigation, IROS, 2004.

[3] B. Williams, M. Cummins, J. Neira, P. Newman, I. Reid, and J. Tardós, A comparison of loop closing techniques in monocular SLAM, Robot. Auton. Syst., 2009.

[4] Y. Taguchi, and Y.D. Jian, S. Ramalingam, and C. Feng, Point-Plane SLAM for Hand-Held 3D Sensors, ICRA, 2013.

[5] A.J.B. Trevor, J.G. Rogers, and H.I. Christensen, Planar surface SLAM with 3D and 2D sensors, ICRA, 2012.

[6] Y. Bok, Y. Jeong, D.G. Choi, and I.S. Kweon, Capturing Village-level Heritages with a Hand-held Camera-Laser Fusion Sensor, Int. J. Comput. Vision, 2011.

[7] D. Nister, A minimal solution to the generalised 3-point pose problem, CVPR, 2004.

[8] M. Lhuillier, Incremental Fusion of Structure-from-Motion and GPS Using Constrained Bundle Adjustments, IEEE Trans. Pattern Anal. Mach. Intell., 2012.

[9] A. Chiuso, P. Favaro, H. Jin, and S. Soatto, 3-D motion and structure from 2-D motion causally integrated over time: Implementation, ECCV, 2000.

[10] D. Nister, O. Naroditsky, and J. Bergen, Visual odometry, CVPR, 2004.

[11] R. Koch, Dynamic 3-d scene analysis through synthesis feedback control, IEEE Trans. Pattern Anal. Mach. Intell., 1993.

[12] A. Comport, E. Malis, and P. Rives, Accurate Quadri-focal Tracking for Robust 3D Visual Odometry, ICRA, 2007.

[13] P.J. Besl, and N.D. McKay, A method for registration of 3-D shapes, IEEE Trans. Pattern Anal. Mach. Intell., 1992.

[14] S. Rusinkiewicz, and M. Levoy, Efficient variants of the ICP algorithm, 3DIM, 2001.

[15] C. Kerl, J. Sturm, and D. Cremers, Dense Visual SLAM for RGB-D Cameras, IROS, 2013.

[16] P. Henry, M. Krainin, E. Herbst, X. Ren, and D. Fox, RGB-D Mapping: Using Kinect-Style Depth Cameras for Dense 3D Modeling of Indoor Environments, IJRR, 2012.

[17] R.A. Newcombe, S. Izadi, O. Hilliges, D. Molyneaux, D. Kim, A.J. Davison, P. Kohli, J. Shotton, S. Hodges, and A. Fitzgibbon, KinectFusion: Real-time Dense Surface Mapping and Tracking, ISMAR, 2011.

[18] A. Nüchter, K. Lingemann, J. Hertzberg, and H. Surmann, 6D SLAM - 3D Mapping Outdoor Environments: Research Articles, J. Field Robot., 2007.

[19] M. Tamaazousti, V. Gay-Bellile, S.N. Collette, S. Bourgeois, and M. Dhome, NonLinear refinement of structure from motion reconstruction by taking advantage of a partial knowledge of the environment, CVPR, 2011.

[20] A. Fitzgibbon, Robust registration of 2D and 3D point sets, Image and Vision Computing, 2003.

[21] B. Triggs, P. Mclauchlan, R. Hartley, and A. Fitzgibbon, Bundle adjustment a modern synthesis, Vision Algorithms: Theory and Practice, LNCS, 2000.

[22] A. Geiger, P. Lenz, C. Stiller, and R. Urtasun, Vision meets Robotics: The KITTI Dataset, International Journal of Robotics Research, 2013.

[23] D. P. Paudel, C. Demonceaux, A. Habed, and P. Vasseur, Localization of 2D Cameras in a Known Environment using Direct 2D-3D Registration, ICPR, 2014.