Localization of 2D Cameras in a Known Environment using Direct 2D-3D Registration

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

Localization of 2D Cameras in a Known Environment using Direct 2D-3D Registration

Danda Pani Paudel Cédric Demonceaux
Le2i UMR 6306
CNRS, University of Burgundy, France
Danda-pani.Paudel@u-bourgogne.fr
Cedric.Demonceaux@u-bourgogne.fr

Adlane Habed
ICube UMR 7357
CNRS, University of Strasbourg, France
Adlane.Habed@icube.unistra.fr

Pascal Vasseur
LITIS EA 4108
University of Rouen, France
Pascal.Vasseur@insa-rouen.fr

***Abstract*—In this paper we propose a robust and direct 2D-to-3D registration method for localizing 2D cameras in a known 3D environment. Although the 3D environment is known, localizing the cameras remains a challenging problem that is particularly undermined by the unknown 2D-3D correspondences, outliers, scale ambiguities and occlusions. Once the cameras are localized, the Structure-from-Motion reconstruction obtained from image correspondences is refined by means of a constrained nonlinear optimization that benefits from the knowledge of the scene. We also propose a common optimization framework for both localization and refinement steps in which projection errors in one view are minimized while preserving the existing relationships between images. The problem of occlusion and that of missing scene parts are handled by employing a scale histogram while the effect of data inaccuracies is minimized using an M-estimator-based technique.**

I. Introduction

Structure-from-Motion (SfM) methods reconstruct an unknown 3D scene from a set of correspondences in two or more views. Such methods customarily refine the pose and reconstruction by minimizing the re-projection error in all the views (i.e. through Bundle Adjustment (BA)). For a better camera localization, it is highly desirable to benefit from the knowledge of the scene when available. Besides scene augmentation and 2D-to-3D data fusion, camera localization in a known 3D scene has the potential of playing a key role in collaborative 3D reconstruction from networks of moving cameras. Furthermore, even for a single camera, when a large sequence is captured, it can be more beneficial to refine the pose of the next frame using a previously acquired reliable 3D rather than performing BA for every new frame.

The 2D-to-3D registration problem is approached in the literature through direct and indirect methods. The direct registration methods rely on establishing feature correspondences such as points, lines, planes, skylines and building bounding boxes between the images and the 3D scene. The point-based matching methods proposed in [11], [6] require the 3D scene along with a scale invariant feature descriptor (SIFT) for each point. Correspondences are obtained by matching these feature descriptors to that of image feature points. Establishing reliable correspondences may be undermined by the absence of such descriptors in the provided scene points as well as by the variability of the illumination conditions during the 2D and 3D acquisitions. Methods relying on higher level features, such as lines [1], planes [12] and building bounding boxes [7], are generally suitable for Manhattan World scenes (or the like) and hence applicable only in such environments. Skylines-based methods [10] as well as methods relying on a predefined 3D model [2] are, likewise, of limited applicability. Indirect methods are performed either by 3D-to-3D registration or by finding some appropriate registration parameters. Methods based on 3D-to-3D registration are performed using the (rigid or non-rigid) Iterative Closest Point (ICP) algorithm between SfM reconstruction and the known scene. However, such registration is not straightforward due to the unknown scale of reconstruction. For instance, this scale ambiguity is handled by an extension of the 4-point congruent sets algorithm in [3]. On the other hand, registration based on complex parameters, such as mutual information [15] and region segmentation [13], are based on single images. Therefore, each camera requires its own initialization and is individually localized independently from the rest of the cameras. Cameras that are localized in this fashion may fail to satisfy the multiview geometric constraints (such as the epipolar constraint in two images). Regarding the camera pose refinement, the method proposed in [12] provides a very good insight into the way the camera pose can be improved when a partial 3D is known. However, this method uses only information about scene planes and assumes that the initial 2D-to-3D registration has already been carried out

In this paper, we propose a method for direct 2D-to-3D registration of multiple calibrated cameras and a known scene. We also propose a constrained nonlinear optimization framework that takes advantage of the knowledge of the scene to simultaneously refine the pose of all cameras once the coarse registration is obtained. Our method demands only a rough knowledge of the pose of only one of the cameras and, apart from 3D scene point coordinates, requires no other knowledge regarding the geometry of the input scene. We assume that the point correspondences across images are available but 2D-3D correspondences are unknown. To our knowledge, there is no method that makes use of both 2D and 3D information without 2D-3D correspondences. Note that methods such as BA with known scene [14] and PnP [4] require such 2D-3D correspondences to be established. In practice, good 2D correspondences between instantaneously captured images can be obtained by using state-of-the-art feature descriptors (such as SIFT). Starting from a roughly known pose of one camera, registration is carried out by minimizing the projection error in one view while preserving the relationship between all pairs of images in the sequence. The 2D-to-3D correspondences required here are selected such that every pair of corresponding

points in two images yields a 3D counterpart whose distance to the scene is minimal while all 3D points emanating from those images share a common relative scale. This distance measurement is derived from epipolar geometry and hence independent from the relative scale. Using such distance measurement allows to avoid the scale problem that arises during the reconstruction. The true relative scale can then be recovered by building a scale histogram where 2D-to-3D correspondences vote for their relative scales. The 3D scene required for our method is no more than a set of points, which in practice, can easily be acquired from other 2D cameras or 3D sensors. Likewise, a rough pose estimation of one camera can be obtained either from SfM itself or from odometry. This pose refinement process also minimizes the same objective used for 2D-to-3D registration while additionally enforcing the epipolar constraint between pairs of views. Both registration and refinement are incorporated in a common optimization framework whose optimal solution is obtained by an iterative method. We use an M-estimator-based iterative weighting scheme for cost as well as the constraints to reduce the effect of inaccuracies in the data. The scale histogram we have built to find the correspondences and the true scale can efficiently detect the occluded/missing parts of the scene. The registration part 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. Our experiments show that the accuracy of our refinement method is significantly better than that of the commonly used BA.

Our paper is organized as follows: Section II introduces the necessary background and notations. In Section III, we formulate our optimization problem for the 2D-to-3D registration problem and propose an algorithm for solving it. Experiments using both synthetic and real data are presented in Section IV. Finally, Section V concludes our work.

II. Notation and background

Let $X^i, i = 1 \dots n$ be the points from a known scene in 3-space represented in a world co-ordinate system O_w. Our goal is to accurately localize a set of p calibrated cameras with respective camera co-ordinate systems $O_1, O_2 \dots O_p$ with respect to O_w by taking advantage of the known scene. Let R and t be the rotation and translation, commonly known as "pose", of the first camera with respect to O_w. If x_1^j and $x_2^j, j = 1 \dots m$ are the corresponding set of feature points in two views, the relative pose of the second camera with respect to the first one (R', t') can be obtained by decomposing the essential matrix using [9]. Note that, we use the term "relative pose" for the transformation from one camera co-ordinate system to another. Let $\widetilde{X}^j, j = 1 \dots m$ be the reconstruction from two views in O_1. Every rotation matrix R is represented by a 4×1 vector of quaternions q unless mentioned otherwise. Both 3D and 2D points are represented by 3-vectors, the latter being in the homogeneous representation. The 2D-to-3D correspondences are specified by a function ϕ. For instance, we denote by $\phi(j)$ the function that maps each pair of 2D points $x_1^j \leftrightarrow x_2^j$ to the corresponding 3D point X^i. The distance between two rotation matrices is measured by computing the spectral norm of their difference. For a matrix A, its spectral norm is denoted as $|||A|||$. Two given up-to-scale translation vectors are compared by measuring the angle between them.

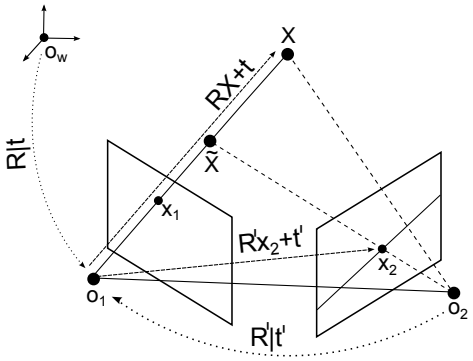

Fig. 1: Triangulation.

III. 2D-to-3D registration method

In this section, we establish the relationships between pairs of image points in two views and the known scene. We propose an optimization framework using these relationships whose optimal solution is the required registration parameters. We also present and discuss an algorithm for solving this optimization problem.

A. Problem formulation

The relationship between 2D and 3D points is depicted in the triangulation diagram given in Fig. 1. The inner product between the normal of the plane $[t']_\times R' x_2$ and the vector $RX + t$ lying on the same plane results in the relationship

$$f(R, t, R', t') = (RX + t)^T [t']_\times R' x_2 = 0. \tag{1}$$

Since the vector x_1 should align with the vector $RX + t$,

$$RX + t = \alpha x_1 \tag{2}$$

must also be satisfied for some unknown scale factor α. The scale factor can be eliminated by using cross-product thus leading to

$$g(R, t) = |||[x_1]_\times (RX + t)||^2 = 0. \tag{3}$$

Furthermore, the epipolar constraint between two views is expressed as

$$h(R', t') = x_1^T [t']_\times R' x_2 = 0. \tag{4}$$

While (3) locates the first camera, (1) locates the second camera with respect to the world frame while preserving its relationship to the first one. Similarly, (4) localizes the second camera with respect to the first one. Equations (1), (3) and (4) are obviously redundant. However, in the presence of noisy data and unknown correspondences all constraints must be enforced: satisfying only the non-redundant conditions does not necessarily satisfy all of them. In addition, (4) makes use of the unknown part of the scene as well. Therefore, all three equations will be incorporated in our optimization framework in which (1) is chosen to be the objective (as it includes the pose of both the cameras) while the rest of the constraints are used as constraints.

Let us consider for now the problem of localizing two cameras given known 2D-to-2D ($x_1^j \leftrightarrow x_2^j$) and unknown 2D-to-3D ($x_1^j \leftrightarrow x_2^j \leftrightarrow X^{\phi(j)}$) correspondences in a noisy

environment. The 2D-to-3D registration problem then boils down to finding the optimal ϕ through optimization. Finding the optimal values of R, t, R' and t', once ϕ is obtained, will be referred to as the camera pose refinement. Stating both the registration and refinement problems in a common optimization framework can be written as

$$\begin{aligned} \underset{q,t,q',t',\phi}{\text{minimize}} \quad & \sum_{j=1}^{m}\{(RX^{\phi(j)}+t)^T[t']_\times R'x_2^j\}^2 \\ \text{subject to} \quad & ||[x_1^j]_\times(RX^{\phi(j)}+t)||^2=0, \\ & \{(x_1^j)^T[t']_\times R'x_2^j\}^2=0,\ j=1\dots m \\ & ||q||^2=1,\ ||q'||^2=1,\ ||t'||^2=1. \end{aligned} \tag{5}$$

The optimization problem (5) 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 first camera center through an image point. All such points satisfy the epipolar constraint and hence lead to correspondence ambiguities, and (b) extra 2D or missing 3D points resulting in invalid 2D-to-3D correspondences. We address both of these problems by assigning the weights derived from a scale histogram for each of these correspondences. For 3D-to-3D correspondences $\widetilde{X}^j \leftrightarrow X^{\phi(j)}, j=1\dots m$, the relative scale of the reconstruction is

$$s(j)=\frac{||\widetilde{X}^j||}{||RX^{\phi(j)}+t||}, \quad j=1\dots m. \tag{6}$$

Since all the points undergo the same scale change during the reconstruction, for the ideal case $s(j)=s(i)\,\forall\{i,j\}\epsilon 1\dots m$. In practice, when the histogram $H(u)$, $u=1\dots b$ of these scales is built, it holds the highest number of samples in the bin corresponding to the true scale. If u_{max} is the bin with highest number of samples, then the weights are distributed as

$$w(j)=\begin{cases}1 & s(j)\epsilon H(u_{max}) \\ 0 & otherwise.\end{cases} \tag{7}$$

Furthermore, the effect of the inaccuracies in the data is reduced by introducing a robust estimation technique. Hence, the optimization problem (5), after including the robust estimation and histogram-based weighting, can be re-written as

$$\begin{aligned} \underset{q,t,q',t',\phi}{\text{minimize}} \quad & \sum_{j=1}^{m} w(j)\rho((RX^{\phi(j)}+t)^T[t']_\times R'x_2^j) \\ \text{subject to} \quad & w(j)\rho(||[x_1^j]_\times(RX^{\phi(j)}+t)||)=0, \\ & \rho((x_1^j)^T[t']_\times R'x_2^j)=0,\ j=1\dots m \\ & |||q||^2=1, ||q'||^2=1, |t'||^2=1. \end{aligned} \tag{8}$$

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

$$\rho(y)=\begin{cases}\frac{y^6}{6}-\frac{\xi^2y^4}{2}+\frac{\xi^4y^2}{2} & for\,|y|<\xi \\ \frac{\xi^6}{6} & otherwise\end{cases} \tag{9}$$

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

Note that, the cost and first constraint functions consider only the known part of the scene. However, the second constraint includes the unknown part of the scene as well. The optimal registration parameters are obtained by iteratively solving this optimization problem. Each iteration breaks down the problem into two: (a) 2D-to-3D registration and (b) Camera pose refinement.

B. 2D-to-3D registration

A camera pair is localized in the scene by iteratively estimating the registration parameters R, t and ϕ. This is performed by solving

$$\begin{aligned} \underset{R,t,\phi}{\text{minimize}} \quad & \sum_{j=1}^{m} w(j)\{(RX^{\phi(j)}+t)^T[t']_\times R'x_2^j\}^2 \\ \text{subject to} \quad & w(j)|||[x_1^j]_\times(RX^{\phi(j)}+t)||^2=0, j=1\dots m. \end{aligned} \tag{10}$$

In general, finding ϕ is also a part of the optimization process. In this case, our choice of ϕ is such that it maps every pair of image points to a 3D point that minimizes the distance between them. This is written as

$$\phi(j)=\underset{i=1\dots n}{\text{argmin}}\ d(R,t,X^i,x_1^j,x_2^j),\ j=1\dots m, \tag{11}$$

where $d(R,t,X,x_1,x_2)$ is a distance that measures the sum of square of projection errors in two views given by

$$\begin{aligned} d(R,t,X,x_1,x_2)=&||[x_1]_\times(RX+t)||^2 \\ &+((RX+t)^T[t']_\times R'x_2)^2. \end{aligned} \tag{12}$$

Hence, the optimal pose of the first camera is

$$\begin{aligned} &\{R^*,t^*\}= \\ &\underset{R,t}{\text{argmin}} \sum_{j=1}^{m} w(j)\{(RX^{\phi(j)}+t)^T[t']_\times R'x_2^j\}^2 \\ &\text{subject to} \quad w(j)|||[x_1^j]_\times(RX^{\phi(j)}+t)||^2=0, j=1\dots m. \end{aligned} \tag{13}$$

Since both cost and constraint functions are linear in R and t, the solution to this problem can be obtained by singular value decomposition. Note that, the linear solution in this case does not consider the rotation matrices in their quaternionic representation. The obtained solution is forced to be a rotation matrix and then converted to quaternions.

C. Camera pose refinement

Coarse registration obtained from 2D-to-3D registration is refined using a constrained nonlinear optimization process. This step refines the pose of the first camera as well as the relative pose of the second camera using the knowledge of 3D scene. Once, the correspondence function ϕ is known, the registration parameters are refined by solving the following optimization problem

$$\begin{aligned} &\{q^*,t^*,q'^*,t'^*\}= \\ &\underset{q,t,q',t'}{\text{argmin}} \sum_{j=1}^{m} w(j)\rho((R_qX^{\phi(j)}+t)^T[t']_\times R_{q'}x_2^j) \\ &\text{subject to} \quad w(j)\rho(||[x_1^j]_\times(R_qX^{\phi(j)}+t)||)=0, \\ &\qquad \rho((x_1^j)^T[t']_\times R_{q'}x_2^j)=0,\ j=1\dots m \\ &\qquad ||q||^2=1, ||q'||^2=1, ||t'||^2=1. \end{aligned} \tag{14}$$

This 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 the interior-point method to solve the constrained nonlinear least-squares problem.

D. The algorithm

Starting from the initial estimate of the parameters $\{R_0, t_0, R'_0, t'_0\}$, obtained from roughly known first camera pose and the essential matrix decomposition, the algorithm iteratively estimates the parameters $\{R_k, t_k, R'_k, t'_k, \phi_k\}$ such that the cost function (8) reduces over the iterations $k = 0 \dots s$ while satisfying its constraints. Each iteration performs the following two steps:

1) **Camera pair alignment**: the camera pair is iteratively aligned with the 3D scene starting from the initial estimate $\{R_{k,0}, t_{k,0}\} = \{R_k, t_k\}$. Each iteration ($l = 0 \dots r$) of this part of the algorithm comprises the following
   a) compute 2D-3D correspondences using (11);
   b) build the scale histogram and compute weights $w(j), j = 1 \dots n$;
   c) update the pose of the first camera using (13).
2) **Simultaneous pose refinement**: starting from the initial estimates $\{R_{k,r}, t_{k,r}, R'_k, t'_k, \phi_{k,r}\}$, the poses of both the cameras are refined by solving (14).

E. Normalization and pose recovery

For the sake of numerical stability, the 3D scene points are normalized such that the distance between its centroid to the first camera is approximately equal to 1. If the initial estimate of the first camera pose is $\{R_0, t_0\}$, such normalization corresponds to $X^i_n = (R_0 X^i + t_0)/||t_0||$, $i = 1 \dots n$. After this transformation, R_0 and t_0 simplify to $I_{3\times3}$ and $0_{3\times1}$ respectively. If the optimal registration parameters obtained from the optimization are R^*, t^*, R'^*, and t'^*; R' and t' are updated to R'^* and t'^*, but R and t are updated to R^*R_0 and $R^*t_0 + ||t_0||t^*$. On the other hand, we also normalize the data during robust estimation i.e. y in Equation (9) is scaled with twice of its median value and ξ is set to 1 whenever it is used. The iterations are terminated when the improvement of the pose between two consecutive iterations $k-1$ and k of both the cameras becomes insignificant. The improvements on the rotational and translational components are computed using

$$e_R = |||R_k - R_{k-1}||| \text{ and } e_t = cos^{-1}\left(\frac{t_k^T t_{k-1}}{||t_k||||t_{k-1}||}\right). \quad (15)$$

Improvements on R' and t' are also computed in the same way. The algorithm terminates when $e_R < T_1$, $e_{R'} < T_1$, $e_t < T_2$, and $e_{t'} < T_2$ for some given thresholds T_1 and T_2.

F. Generalization to multiview

In multiview case of p cameras, the registration parameters estimation using (8) for two views has been generalized to

$$\begin{aligned} \underset{\substack{q_l, q'_l, \phi, \\ t_l, t'_l}}{\text{minimize}} \quad & \sum_{l=1}^{p}\sum_{j=1}^{m} w_l(j)\rho((R_l X^{\phi(j)} + t_l)^T [t'_l]_\times R'_l x^j_{l+1}) \\ \text{subject to} \quad & w_l(j)\rho(|||[x^j_l]_\times (R_l X^{\phi(j)} + t_l)||) = 0, \\ & \rho((x^j_l)^T [t'_l]_\times R'_l x^j_{l+1}) = 0,\ j = 1 \dots m \\ & ||q_l||^2 = 1, ||q'_l||^2 = 1, \\ & ||t'_l||^2 = 1,\ l = 1 \dots p-1. \end{aligned} \quad (16)$$

Our solution to this problem is similar to resection-intersection based BA [14]. Since we do not directly refine the 3D points, the first part of the algorithm performs the registration followed by the refinement method. Once a pair of cameras is localized, the refined pose of the second camera is used to initialize the pose of the first camera of the next pair. As both cameras of the next pair are free to move during refinement, the error introduced in previous pairs does not propagate to the next ones. More importantly, roughly known pose of only one camera suffices for the multiview case as well.

IV. Experiments

We tested our method using synthetic and real data. Our results with synthetic data are compared against those of ICP with classical SfM. The results obtained on two benchmark real datasets and one in-house scene are compared against SfM for two views and BA for multiple views.

A. Simulations

We generated a set of 800 random 3D points scattered on the surface of four faces of a $[-10\ 10]^3$ cube. The cameras were placed about 20 ± 2 units away from the origin with randomly generated rotations while roughly looking towards the centroid of the scene. All scene points were projected onto 256×256 images with zero-skew, 100 pix. focal length and an image-centered principal point. The 2D data were obtained by adding various levels of zero-mean Gaussian noise to the pixel coordinates. 400 out of 800 projected points were randomly selected and used to localize the second camera with respect to the first one using classical SfM. During this process, half of the points are rejected to minimize the effect of outliers thus leading to the reconstruction of only 200 points. The same data were used in our method to perform the registration and refinement. We ran 100 tests for each noise level of standard deviation (0 to 2.0 with a 0.25 step). The simulation results are presented for the two-view case only.

The roughly known R was generated by introducing an error of $[0.05\ 0.075]^c$ in roll, pitch and tilt each. We introduced these relatively small errors in R to observe the improvements when the iterative scheme converges. Similarly, a small error of $\pm 5\%$ is introduced in each translation axis. Nevertheless, these errors are very significant since the scene is relatively far from the cameras. The histogram was built with auto adjustable 10 bins after discarding the scales of less than 0.1 and greater than twice its median. First, we obtained the best possible R, t, R', and t' using classical SfM and ICP. As ICP cannot

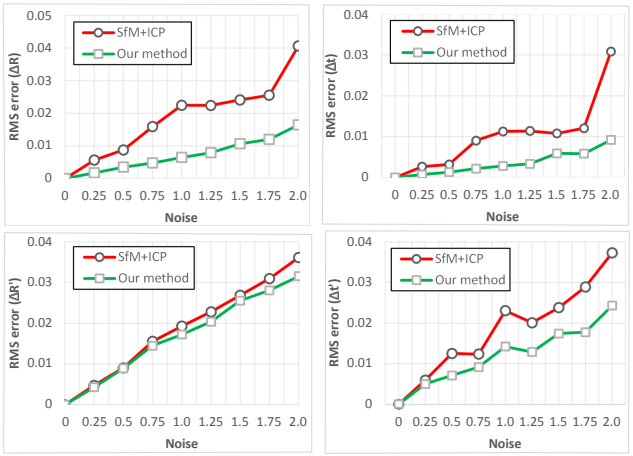

Fig. 2: SfM+ICP vs. Our method with noise; ΔR (left-top), Δt (right-top), $\Delta R'$ (left-bottom), and $\Delta t'$ (right-bottom).

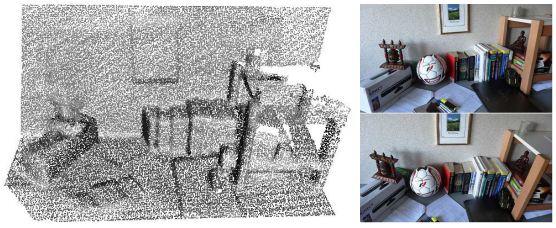

Fig. 3: Left: Kinect 3D scene; Right: image pair.

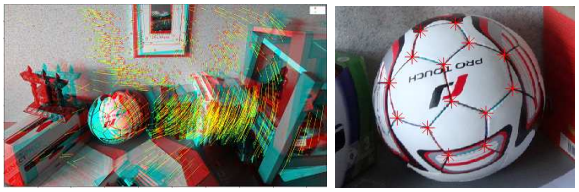

Fig. 4: Left: Correspondences; Right: feature points.

be performed without the knowledge of relative scale, the extra information of scale is recovered with the assumption of the image-based reconstruction being spread all over the provided 3D scene. Note that, our method does not require this extra information of scale. To analyze the improvements on camera pose, we computed the deviation of these results from their ground truth values. The errors ΔR, Δt, $\Delta R'$, and $\Delta t'$ correspond to the residuals computed as in (15). Fig. 2 shows the Root-Mean Square (RMS) plots of the computed errors for various levels of noise. It can be seen that our method performs significantly better than SfM with ICP even when the ICP is favored with extra information of scale. ICP was performed using [5] and the optimization was carried out using the interior point method (MATLAB-R2012a).

B. Real data

For the first experiment with real data, we built the prior 3D scene by registering multiple frames acquired from a 3D sensor (Kinect). This scene was then down-sampled to about 50,000 points as shown in Fig. 3 (left). After the 3D scene is acquired, a standard sized football was placed in the same scene and two 1080×1920 images were captured by a moving camera. These images and their 1198 correspondences are shown in Fig. 3 and Fig. 4. 14 manually selected points from the corners of the Truncated Icosahedron (TI) (Fig. 4 (right)) were retained for assessing the quality of the reconstruction. To overcome the problem of initialization, the first views of both 2D and 3D cameras are captured approximately from the same location while facing towards the same part of the scene.

The final metric reconstruction of the scene is upgraded to Euclidean for the measured length of polygon sides equal to 4.5 cm. Reconstructed TI from two views is placed in the given 3D scene and shown in Fig. 5. We have approximated the circumference of the football by fitting a sphere passing through the vertices of the reconstructed TI. For a quantitative analysis, the following geometric parameters of reconstructed TI are computed: (i) LS: RMS error of the length of sides. (ii) AH: RMS error of the internal angles of hexagons. (iii) AP: RMS error of the internal angles of pentagons. (iv) A-HP: RMS error of Dihedral angles between hexagons and the pentagons. (v) A-HH: Dihedral angle between two hexagons (expected: 138.19). (vi) CS: Circumference of the sphere (expected: 68-70 cm). Table I compares these parameters against FIFA's standard. This is an example of 2D-to-3D data fusion where the reconstruction from two views is added to the 3D scene. This example also demonstrates the handling of occlusion problem because of the football placed in the scene after the 3D acquisition. Furthermore, even when the 3D data is not very accurate, like in this case, it shows that our method still benefits from the scene information. We also tested our method with the public datasets Fountain-P11 and Herz-Jesu-K7 (Fig. 6 from http://cvlabwww.epfl.ch/~strecha). These datasets consist, respectively, of 11 and 7 images of size 3072×2048 along with ground truth partial 3D point clouds of the scenes. To validate the ground truth, the texture was mapped on the scene by back-projecting images using their ground truth projection matrices. Fig. 7 shows that the provided camera poses are very satisfactory (unlike M. Corsini et al. reported in [3]). First, the 3D reconstructions for every consecutive pair of images are obtained using classical SfM. All these results are then refined separately using our method. Results before and after the refinement are compared against the ground truth in Table II.

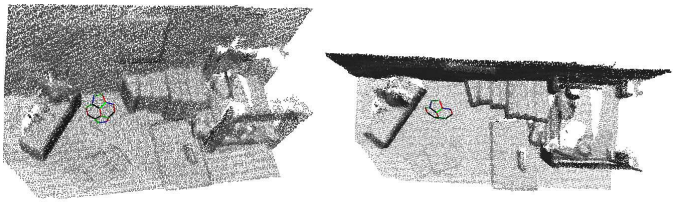

Fig. 5: Two views of the 3D scene with TI.

	LS (cm)	AP	AH	A-HP	A-HH	CS (cm)
SfM	0.201	4.267	2.008	6.195	140.19	76.25
Our method	0.117	2.943	0.863	3.342	139.20	73.10

TABLE I: Geometric parameters.

Fig. 6: Left: Fountain-P11; Right: Herz-Jesu-K7.

Fig. 7: Texture mapping of Herz-Jesu-K7.

The 3D errors shown here are the mean 3D RMS error of all the pairs. During the implementation, we have decimated the 3D scenes to about 50,000 points by uniform down-sampling for a faster computation. About 2000-3000 feature points were selected in each pair of views for the reconstruction. For the multiview case, reconstructions from each consecutive pair of views are registered. Such registration undergoes error accumulation and scale factor drift. We separately refined these results using our method and sparse BA [8]. The results using our method were found to be significantly better than those of BA. We also considered refining our results using BA. Results obtained from BA, our method, and BA performed to refine our results are shown in Table III. It is observed that the BA performed on our results diverges from the ground truth instead of further refinement. Since BA takes only the image information into account and cannot incorporate the 3D knowledge; noise present in the image might be the reason for this divergence. For qualitative analysis, results obtained from BA as well as our method were used to map the texture (Fig. 8). Texture mapping using BA contains many artifacts the most visible of which has been circled in this figure. Note that, as the scene being relatively far from the cameras, even a small error in pose can significantly affect the texture mapping. It clearly shows the pose refinement using our method is very accurate and visually no different from the ground truth.

V. Conclusion

In this paper, we have proposed an optimization framework to accurately localize two or more cameras in a known environment. We have demonstrated the possibility of precisely registering 2D images to 3D scene using only the feature points. Usage of a known 3D scene to refine the camera pose is key to achieve such accuracy. To make it possible, a direct 2D-to-3D registration method has also been integrated in the optimization process.

	Method	Fountain	Herz-Jesu
$\Delta R'$(RMS)	SfM	0.0044	0.0072
	Our method	8.49e-4	0.0013
$\Delta t'$(RMS)	SfM	0.0404	0.0757
	Our method	0.0031	0.0052
3D error	SfM	0.0011	0.0025
	Our method	5.95e-4	0.0018

TABLE II: SfM vs. our method (two views).

	Method	Fountain	Herz-Jesu
ΔR(RMS)	BA	0.0436	0.0123
	Our method	0.0020	0.0067
	Refined	0.0251	0.0080
Δt(RMS)	BA	0.0311	0.0402
	Our method	0.0019	0.0224
	Refined	0.0172	0.0241
3D error	BA	0.0020	0.0069
	Our method	0.0015	0.0068
	Refined	0.0020	0.0069

TABLE III: BA vs. Our method and unsuccessful refinement of our results using BA (multiview).

Fig. 8: Texture mapping: BA (left) and Our method (right).

References

[1] S. Christy and R. Horaud. Iterative pose computation from line correspondences. In *Comput. Vis. Image Underst.*, pages 137–144, January 1999.

[2] M. J. Clarkson, D. Rueckert, D. L. Hill, and D. J. Hawkes. Using photo-consistency to register 2d optical images of the human face to a 3d surface model. In *IEEE Trans. Pattern Anal. Mach. Intell.*, pages 1266–1280, November 2001.

[3] M. Corsini, M. Dellepiane, F. Ganovelli, R. Gherardi, A. Fusiello, and R. Scopigno. Fully automatic registration of image sets on approximate geometry. In *Int. J. Comput. Vision*, pages 91–111, March 2013.

[4] J. A. Hesch and S. I. Roumeliotis. A direct least-squares (DLS) method for PnP. In *ICCV*, 2011.

[5] M. Kjer and J. Wilm. Iterative closest point, 2012.

[6] J. Knopp, J. Sivic, and T. Pajdla. Avoiding confusing features in place recognition. In *ECCV*, 2010.

[7] L. Liu and I. Stamos. Automatic 3d to 2d registration for the photorealistic rendering of urban scenes. In *CVPR*, 2005.

[8] M. A. Lourakis and A. Argyros. SBA: A software package for generic sparse bundle adjustment. In *ACM Trans. Math. Software*, pages 1–30, 2009.

[9] D. Nistér. An efficient solution to the five-point relative pose problem. In *IEEE Trans. Pattern Anal. Mach. Intell.*, pages 756–777, June 2004.

[10] S. Ramalingam, S. Bouaziz, P. Sturm, and M. Brand. Geolocalization using skylines from omni-images. In *ICCV Workshops*, 2009.

[11] T. Sattler, B. Leibe, and L. Kobbelt. Fast image-based localization using direct 2d-to-3d matching. In *ICCV*, 2011.

[12] 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. In *CVPR*, 2011.

[13] A. Taneja, L. Ballan, and M. Pollefeys. Registration of spherical panoramic images with cadastral 3d models. In *3DIMPVT*, 2012.

[14] B. Triggs, P. Mclauchlan, R. Hartley, and A. Fitzgibbon. Bundle adjustment a modern synthesis. In *Vision Algorithms: Theory and Practice, LNCS*, pages 298–375. Springer Verlag, 2000.

[15] P. Viola and W. M. Wells, III. Alignment by maximization of mutual information. In *Int. J. Comput. Vision*, pages 137–154, September 1997.