NonLinear Refinement of Structure from Motion Reconstruction by Taking Advantage of a Partial Knowledge of the Environment

Mohamed Tamaazousti[1] Vincent Gay-Bellile[1] Sylvie Naudet Collette[1] Steve Bourgeois[1] Michel Dhome[2]
[1]CEA LIST, Vision and Content Engineering Lab, Point Courrier 94, Gif-sur-Yvette, F-91191 France
name.Surname@cea.fr

[2]LASMEA (CNRS / UBP) 24 Avenue des Landais, 63177 Aubière Cedex, France
name.Surname@lasmea.univ-bpclermont.fr

Abstract

We address the challenging issue of camera localization in a partially known environment, i.e. for which a geometric 3D model that covers only a part of the observed scene is available. When this scene is static, both known and unknown parts of the environment provide constraints on the camera motion. This paper proposes a nonlinear refinement process of an initial SfM reconstruction that takes advantage of these two types of constraints. Compare to those that exploit only the model constraints i.e. the known part of the scene, including the unknown part of the environment in the optimization process yields a faster, more accurate and robust refinement. It also presents a much larger convergence basin. This paper will demonstrate these statements on varied synthetic and real sequences for both 3D object tracking and outdoor localization applications.

1. Introduction

The pose estimation of a moving camera with respect to an object of interest[1] is an active research topic. Most of the existing solutions consider a camera navigating in a completely known[2] or completely unknown environment.

Model-based tracking solutions exploit the *a priori* knowledge of the object geometry to estimate the relative pose of the camera wrt this object. Usually, this pose is estimated in real-time by matching 3D features extracted from the model with their corresponding observations in the current image [6]. This process implies the visibility of the object of interest during the whole sequence and thus is generally not suited for large environment. On the other hand, Structure from Motion (SfM) and Simultaneous Localization And Mapping (SLAM) solutions estimate the relative motion of a camera without any prior on the scene geometry. These solutions exploit the multi-view relationships to estimate the motion of the camera and eventually reconstruct a 3D map (*i.e.* a sparse 3D points cloud) of the environment. Offline [12, 14] and online [5, 10, 11] solutions have been introduced. These techniques are well suited for large areas since the whole environment is exploited to estimate the camera motion. Nevertheless, monocular solutions are subject to error accumulation and scale factor drift that reduce their application domains.

Recently, some solutions tried to combine model-based and SfM techniques to accurately estimate the pose of a camera in a partially known environment. Bleser *et al*. in [2] exploit the geometric model constraints to initialize (coordinate frame and scale) the map of a SLAM algorithm. Then they switch to a "classic" SLAM process that no longer take the 3D model informations into account. Their algorithm is thus still subject to initialization inaccuracies, error accumulation and scale factor drift, especially in large environments. To solve these problems, Lothe *et al*. in [7] introduced a two step process. First, a standard SfM reconstruction of the whole environment is roughly aligned on the model through nonrigid ICP. Then, a refinement process, that combine both the multi-view geometry relationships and the geometric constraints of a coarse 3D model is used to refine the reconstruction. Their refinement process does not take into account the unknown parts of the environment: the multi-view constraints relative to the unknown environment are no longer guaranteed and cameras that observe few or no parts of the known environment are under-constrained.

In this paper, we introduce an original solution for camera localization in a partially known environment. It combines both geometric informations provided by known and unknown parts of the environment in a nonlinear refinement algorithm similar to the one introduced by Lothe *et al*. [7]. It yields a more accurate and robust refinement that improves

[1]It can be the whole environment.
[2]By known we mean that a 3D model of the observed scene is available.

the reconstruction of the whole environment. We evaluate our approach on synthetic and real data for both 3D object tracking and vehicle localization applications.

Plan. In Section 2, we introduce equations used to refine an initial reconstruction of a known or an unknown environment. Then, we propose different solutions to combine these two types of constraints in a single consistent framework. They are exposed in Section 3 and evaluated in Section 4. In Section 5, we apply the proposed solution for 3D object tracking and vehicle localization applications. Finally, we give our conclusions and discuss future work in Section 6.

Notation. Matrices are designated by sans-serif fonts such as M. Vectors are typeset using italic fonts and expressed in homogeneous coordinates, e.g. $\mathbf{q} \sim (x, y, w)^\mathsf{T}$ where $^\mathsf{T}$ is the transposition and $\sim$ the equality up to a non-zero scale factor. In the following we assume that an initial reconstruction of the observed scene has already been estimated with any monocular SfM algorithms such as [5, 10, 14]. This resulting reconstruction is composed by N 3D points $\{\mathbf{Q}_i\}_{i=1}^N$ and m cameras $\{\mathsf{C}_k\}_{k=1}^m$. We note $\mathbf{q}_{i,k}$ the observation of the 3D point $\mathbf{Q}_i$ in the camera C_k and $\mathcal{A}_i$ the set of camera indices observing $\mathbf{Q}_i$. The projection matrix P_k associated with the camera C_k is given by $\mathsf{P}_k = \mathsf{K}\mathsf{R}_k^\mathsf{T}(\mathcal{I}_3 | -\mathbf{t}_k)$, where K is the matrix of the intrinsic parameters and $(\mathsf{R}_k, \mathbf{t}_k)$ the extrinsic ones. We suppose that we dispose of a purely geometric 3D model of a part of the observed scene. It is composed by a set of planes π. The registration between the world coordinate frame and the 3D model coordinate frame is assumed approximatively known, see Section 5 for details.

2. Nonlinear Refinement of SfM Reconstruction

Refinement of SfM reconstruction is usually based on nonlinear minimization of a cost function. In the following, we describe cost functions used when the environment is unknown (Section 2.1) or known (Section 2.2).

2.1. Nonlinear Refinement in Unknown Environments

The gold standard solution to refine an initial reconstruction of an unknown environment is the Bundle Adjustment (BA) technique that simultaneously optimize the camera poses and the scene structure. An uncommon alternative is to refine only the camera poses through epipolar geometry (EG) constraints and then reconstruct a 3D point cloud by triangulation.

2.1.1 Epipolar Geometry (EG)

It defines the geometric relations between the images captured by two cameras $(\mathsf{C}_1, \mathsf{C}_2)$ observing the same scene from distinct positions. The Epipolar Geometry [4] links two observations $(\mathbf{q}_{i,1}, \mathbf{q}_{i,2})$ of a 3D point $\mathbf{Q}_i$ through the Fundamental matrix: $\mathbf{q}_{i,2}^\mathsf{T}\mathsf{F}\mathbf{q}_{i,1} = 0$, where F is a 3×3 matrix of rank 2. This relationship means that any point $\mathbf{q}_{i,2}$ in the second image matching the point $\mathbf{q}_{i,1}$ in the first one must lie on the epipolar line $l = \mathsf{F}\mathbf{q}_{i,1}$. The Fundamental matrix depends directly on the inter-frames camera motion such as $\mathsf{F} = \mathsf{K}^{-\mathsf{T}}[\mathbf{t}_{1\to2}]_\times \mathsf{R}_{1\to2}\mathsf{K}^{-1}$. Thus, an initial pose estimate may be refined by minimizing the following criteria: $\mathcal{E}\left((\mathsf{R}, \mathbf{t})_{1\to2}\right) = \sum_{i=1}^{N} d_l^2(\mathbf{q}_{i,2}, \mathsf{F}_{2,1}\mathbf{q}_{i,1}) + d_l^2(\mathbf{q}_{i,1}, \mathsf{F}_{1,2}\mathbf{q}_{i,2})$, where $d_l(\mathbf{q}, \mathbf{l})$ is the point-to-line distance between a point $\mathbf{q}$ and a line $\mathbf{l}$ such as $d_l^2(\mathbf{q}, \mathbf{l}) = \frac{(\mathbf{q}^\mathsf{T}\mathbf{l})^2}{\|\mathbf{l}\|^2 w^2}$.

This principle can be extended in the multi-view geometry case. The displacement of the moving camera is then refined by minimizing the following cost function:

$$\mathcal{E}\left(\left\{(\mathsf{R}, \mathbf{t})_{p\to p+1}\right\}_{p=1}^{m-1}\right) = \sum_{i=1}^{N}\sum_{j\in\mathcal{A}_i}\sum_{k\in\mathcal{A}_i}^{k\neq j} d_l^2(\mathbf{q}_{i,j}, \mathsf{F}_{j,k}\mathbf{q}_{i,k}), \tag{1}$$

where, $\mathsf{F}_{j,k}$ is the Fundamental matrix between the image pair (j, k).

2.1.2 Bundle Adjustment (BA)

Bundle Adjustment assigns the problem of simultaneously refining the 3D points describing the scene geometry as well as the camera poses. Bundle Adjustment [16] minimizes the sum of square differences between the projected 3D points and the associated image observations. This geometric distance is called the reprojection error. The optimized parameters are the coordinates of the N 3D points and the six extrinsic parameters of the m camera poses. The total number of parameters is then $3N + 6m$. The BA cost function is given by:

$$\mathcal{E}\left(\{\mathsf{R}_k, \mathbf{t}_k\}_{k=1}^m, \{\mathbf{Q}_i\}_{i=1}^N\right) = \sum_{i=1}^{N}\sum_{j\in\mathcal{A}_i} d^2(\mathbf{q}_{i,j}, \mathsf{P}_j\mathbf{Q}_i), \tag{2}$$

where $d^2(\mathbf{q}, \mathbf{q}') = \|\mathbf{q} - \mathbf{q}'\|^2$ is the point-to-point distance.

2.2. Nonlinear Refinement in Known Environments

In this section, we describe how the additional constraints provided by the 3D model of the whole environment can be exploited. We describe two non-linear cost functions that combine multi-view geometry relationships and planar constraints [1, 15] provided by a 3D model in a

single term. They have in common to minimize a residual error expressed in pixel.

2.2.1 Homography Constraints

In two-view geometry, two images observing a same plane π are linked by an Homography H. Let $\mathbf{q}_{i,1}$ be the observation of a point $\mathbf{Q}_i \in \pi$ in the first view and $\mathbf{q}_{i,2}$ the observation in the second view, then $\mathbf{q}_{i,1} \sim \mathsf{H}\mathbf{q}_{i,2}$. This is the equivalent of the Epipolar Geometry relationship for the planar case. The Homography H induced by the plane π is given by:

$$\mathsf{H} = \mathsf{K}(\mathsf{R} - \frac{\mathbf{t}\mathbf{n}^\mathsf{T}}{d})\mathsf{K}^{-1}, \tag{3}$$

where, $\mathbf{n}$ is the normal of the plane and d the distance between C_1 and the plane. This relationship has been used by Simon *et al*. in [13] to refine an initial SfM reconstruction. The following cost function is minimized:

$$\mathcal{E}\left(\left\{(\mathsf{R},\mathbf{t})_{p\to p+1}\right\}_{p=1}^{m-1}\right) = \sum_{i=1}^{N}\sum_{j\in\mathcal{A}_i}\sum_{k\in\mathcal{A}_i}^{k\neq j} d^2(\mathbf{q}_{i,j}, \mathsf{H}_{j,k}^{\pi_i}\mathbf{q}_{i,k}), \tag{4}$$

where $\mathsf{H}_{j,k}^{\pi_i}$ is the Homography induced by the observation of plane π_i by the cameras j and k. We note that this cost function does not taken into account the 3D points. Only the camera poses are optimized.

2.2.2 Bundle Adjustment with Model Constraints

The cost function described above includes model constraints through the optimization of camera poses. We propose a novel cost function that contrary to the previous one optimizes explicitly the structure. The main idea is that a 3D point $\mathbf{Q}_i$ lying on a plane π_i has only two degrees of freedom. Lets M^{π_i} the transfer matrix between the coordinate frame of plane π_i and the world coordinate frame then $\mathbf{Q}_i = \mathsf{M}^{\pi_i}\mathbf{Q}_i^{\pi_i}$ where $\mathbf{Q}_i^{\pi_i} = (X^{\pi_i}, Y^{\pi_i}, 0, 1)^\mathsf{T}$ and (X^{π_i}, Y^{π_i}) are the coordinates of $\mathbf{Q}_i$ in the plane π_i coordinate frame. This relation can be used to optimize a SfM reconstruction with model constraints by minimizing the following cost function:

$$\mathcal{E}\left(\{\mathsf{R}_j,\mathbf{t}_j\}_{j=1}^m, \{\mathbf{Q}_i^{\pi_i}\}_{i=1}^N\right) = \sum_{i=1}^{N}\sum_{j\in\mathcal{A}_i} d^2(\mathbf{q}_{i,j}, \mathsf{P}_j\mathsf{M}^{\pi_i}\mathbf{Q}_i^{\pi_i}). \tag{5}$$

In practice, the reconstructed 3D points do not exactly belong to the planes of the model. A preliminary step is then required to project each 3D point $\mathbf{Q}_i$ on its associated plane π_i (see Section 3.3 for more details).

3. Nonlinear Refinement in a Partially Known Environment

In the previous section, we have presented several cost functions to refine an initial SfM reconstruction of a known (Eq. (4) and (5)) or an unknown (Eq. (1), (2)) environment. In this section, we will describe how they can be merged in a single nonlinear optimization process that takes both informations provided by the known and the unknown parts of the environment.

A preliminary step is required to decide which 3D point $\mathbf{Q}_i$ of the initial reconstruction belong to the model. This point-to-model associations problem is achieved through ray tracing from the different observations $\{\mathbf{q}_{i,j}\}_{j\in A_i}$ of $\mathbf{Q}_i$. Thus, card(A_i) votes are obtained. When 3D points are assigned to different planes by different cameras (for example 3D points near the boundaries) we take the majority choice. Ones the classification between the known and unknown parts of the environment is done, 3D points that have been associated to one plane π_i of the model have to be projected on it before minimizing Eq. 5 (not required for Eq. 4). We take the barycenter of the intersections between the rays and the plane π_i. We note $\mathcal{M}$ the set of 3D point indices associated to the model and $\mathcal{U}$ the set of remaining 3D point indices that constitute the unknown part of the environment, with card$(\mathcal{M})$ + card$(\mathcal{U}) = N$.

3.1. Robust Estimation

Inaccuracies in the coordinate frame registration and in the SfM reconstruction introduce many wrong point-to-model associations that will fail the optimization process. To deal with those outliers a robust estimation is used through the Geman-McClure M-estimator $\rho(r, c) : \mathbb{R} \to [0 \cdot\cdot 1]$ where

$$\rho(r, c) = \frac{r^2}{r^2 + c^2}, \tag{8}$$

with, r is a residual error of any cost functions described previously (1), (2), (4) or (5), and c is the rejecting threshold. It is automatically estimated with the Median of Absolute Deviation (MAD) such as $c = median(\mathbf{r}) + 1.4826\text{MAD}(\mathbf{r})$ where $\mathbf{r}$ is a vector concatenating the residuals of any cost function. Note that the MAD assumes a normal distribution of the residuals.

3.2. Compound Cost Functions

Combining Eq. (1) or (2) with Eq. (4) or (5) is not obvious even if this different cost functions share the same unit (pixel). In fact, they do not necessarily share the same magnitude: error residuals associated to the known part of the environment have generally higher values. We model the combination of known and unknown parts of the environment as a bi-objective least square problem.

3.2.1 Consistent Combinations

We try to stay consistent in the combination choices and proposed only two compound cost functions. The first one,

$$\mathcal{E}\left(\left\{(\mathsf{R},\mathbf{t})_{p\to p+1}\right\}_{p=1}^{m-1}\right)=\underbrace{\sum_{i\in\mathcal{U}}\sum_{j\in A_i}\sum_{k\in A_i}^{k\neq j}\rho\left(d_l^2(\mathbf{q}_{i,j},\mathsf{F}_{j,k}\mathbf{q}_{i,k}),c_1\right)}_{\text{Unknown part of the environment (E)}}+\underbrace{\sum_{i\in\mathcal{M}}\sum_{j\in\mathcal{A}_i}\sum_{k\in\mathcal{A}_i}^{k\neq j}\rho\left(d^2(\mathbf{q}_{i,j},\mathsf{H}_{j,k}^{\pi_i}\mathbf{q}_{i,k}),c_2\right)}_{\text{Known part of the environment (M)}} \tag{6}$$

$$\mathcal{E}\left(\{\mathsf{R}_j,\mathbf{t}_j\}_{j=1}^m,\{\mathbf{Q}_i\}_{i\in\mathcal{U}},\{\mathbf{Q}_i^{\pi_i}\}_{i\in\mathcal{M}}\right)=\underbrace{\sum_{i\in\mathcal{U}}\sum_{j\in\mathcal{A}_i}\rho\left(d^2(\mathbf{q}_{i,j},\mathsf{P}_j\mathbf{Q}_i),c_1\right)}_{\text{Unknown part of the environment (E)}}+\underbrace{\sum_{i\in\mathcal{M}}\sum_{j\in\mathcal{A}_i}\rho\left(d^2(\mathbf{q}_{i,j},\mathsf{P}_j\mathsf{M}^{\pi_i}\mathbf{Q}_i^{\pi_i}),c_2\right)}_{\text{Known part of the environment (M)}} \tag{7}$$

uses homography constraints (Eq. (4)) for parts of the environment associated to the piecewise planar model and its equivalent for the unknown structure *i.e.* the epipolar geometry (Eq. (1)). They have in common to optimize only the inter-frame cameras displacement. The second compound cost function combines Eq. (2) and (5) that use explicitly the 3D points in a Bundle Adjustment framework. Thus, 3D points associated to the 3D model have only two degrees of freedom whereas 3D points of the unknown part of the environment have three degrees of freedom.

3.2.2 Weighting through Robust Estimation

One challenging issue in bi-objective minimization is to control the influence of each terms. This is usually done through a weighting parameter that is fixed experimentally or via cross validation [3]. We propose a simplest alternative: the influence of each term is directly controlled through the rejecting threshold of the robust estimator. We have seen above that a robust estimator have to be used to deal with wrong point-to-model associations *i.e.* for the cost functions of the known part of the environment. We also apply the robust estimation to the cost functions of the unknown part of the environment since the Geman-McClure M-estimator normalize the residual. The resulting compound cost functions are then given by Eq. (6) and (7).

Thus there is several possibilities to control the influence of each term trough the rejecting threshold. We will explore three of them:

- combination 1: $c_1 = c_{Env}$ and $c_2 = c_{Model}$
- combination 2: $c_1 = c_2 = c_{All}$
- combination 3: $c_1 = c_2 = c_{Model}$

where, c_{Model} is the rejecting threshold estimated on the model-based residuals such as those used in Eq. (4), (5), c_{Env} is the rejecting threshold estimated on the residuals of the unknown parts of the environment such as those used in Eq. (1) or (2) and c_{All} is the rejecting threshold estimated on all residuals.

For the combination 1 there is a rejecting threshold associated to each term whereas for combination 2 and 3 only one threshold is estimated. The difference is that it is evaluated on all the residuals for combination 2 and only on the residuals associated to the known part of the environment for combination 3. Combination 2 considers that those two types of residuals have the same order of magnitude. In opposition, combinations 1 and 3 take into account the fact that residuals associated to the known part of the environment have generally higher values. Combination 1 treats the known and the unknown parts of the environment identically whereas combination 3 will favor the former during the optimization process while guaranteeing that the unknown environment constraints are still verified. This three combinations are evaluated on synthetic data in Section 4.2.

3.3. Iterative Optimization

During the optimization process the point-to-model associations and the rejecting threshold of the robust estimator have to be re-examined to guarantee an optimal convergence. The following steps are thus iterated until convergence:

1. Association of the 3D points $(\mathbf{Q}_i)_{i\in\mathcal{U}\cup\mathcal{M}}$ to the known or the unknown parts of the environment.
2. Projecting the 3D points $(\mathbf{Q}_i)_{i\in\mathcal{M}}$ on their associated plane π_i (this step is only for Eq. (7)).
3. Compute the rejecting thresholds c_1 and c_2.
4. Minimization of (6) or (7) by the Levenberg Marquardt (LM) algorithm [9] (few iterations).
5. Triangulation of 3D points $(\mathbf{Q}_i)_{i\in\mathcal{U}\cup\mathcal{M}}$ for Eq. (6) and $(\mathbf{Q}_i)_{i\in\mathcal{M}}$[3] for Eq. (7) with the estimated camera poses.

4. Evaluation on Synthetic Data

In this section we compare four algorithms on a synthetic sequence generated with a 3D computer graphics software: BA_M, BA_M&E, EG_M and EG_M&E[4]. They minimize Eq. (5), (7), (4) and (6) respectively with the procedure described in Section 3.3. Note that for BA_M and EG_M the Geman Mclure M-estimator (Section 3.1) has also been used to deal with wrong point-to-plane associations. We firstly compare the three combination choices of Section

[3] For Eq. (7), 3D points of the known part of the environment have to be triangulated due to the point-to-model associations re-examination.

[4] M means that only the model constraints, *i.e.* the known part of the environment are used whereas M&E means that both the model constraints and the information provided by the unknown part of the environment are taken into account.

3.2 for the BA_M&E algorithm. Then we compare the four algorithm described above, in terms of convergence basin, convergence speed and accuracy against inaccuracies of SfM reconstruction and coordinate frame registration.

4.1. The Cube Sequence

This sequence, represented in Figure 1 is composed by a main cube over a textured ground with other smaller cubes that partially occult it. The object of interest, *i.e.* for which a 3D model is available, is the main cube. The environment is thus composed by the ground and the small cubes. The camera trajectory is a circle of 3 meters radius around the main cube.

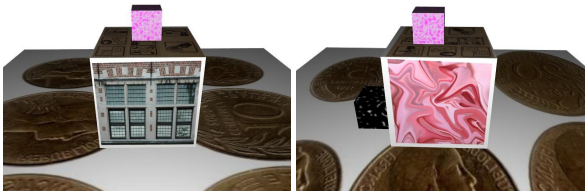

Figure 1. Illustration of the Cube sequence.

4.2. Combination Choices

We compare for the BA_M&E algorithm the three combinations proposed in Section 3.2. The initial SfM reconstruction is obtained with the algorithm described in [10]. It is composed by 21 cameras (keyframes images) and about 1500 3D points. The world coordinate frame and the scale of the reconstruction have been fixed by the ground truth. Then the resulting reconstruction is refined by minimizing one of the three combinations on Eq. (7) with the procedure described in Section 3.3.

Figure 2 (Left) shows the distribution for both types of residuals after the step 2 of the minimization process, see Section 3.3. As expected the magnitudes of the residual errors associated to the known part of the environment are higher than those associated to the unknown one. It explains that the combination 2 presents the worst results as seen in Figure 2 (Right) since its rejecting threshold is underestimated: most of the model-based residuals are then discarded by the robust estimator. Using combination 1 and optimizing only Eq.(5) (*i.e.* the BA_M algorithm) gives similar results. Finally the combination 3 outperforms other solutions. It proves that the known part of the environment has to be favored during the optimization process while guaranteeing that the unknown environment constraints are still verified.

Note that similar results have been obtained on other synthetic sequences and with the EG_M&E algorithm. They are not reported to not overload the paper. In the following we only consider the combination 3 for Eq.(6) and (7).

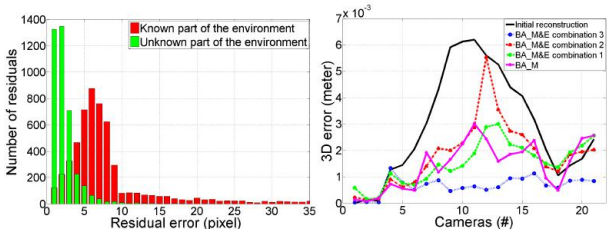

Figure 2. Left: Residual Errors distributions. In green, (resp. in red) the distribution of the residual errors associated to the environment (resp. the 3D model). Right: Cameras position errors expressed in meter for the different combinations.

4.3. Comparison of the Four Algorithms.

Experimental protocol. We compare the contribution of the four algorithms by simulating the different sources of errors *i.e.* coordinate frame registration, drift, *etc*. From the camera poses of the ground truth, a sparse 3D point cloud has been generated by triangulation of matching interest-points along the sequence. Then, two kinds of perturbation have been generated on this initial reconstruction.

- The first test (TEST RIGID) simulates inaccuracy of coordinates frames registration (world / model). A rigid perturbation is applied on the global reconstruction (cameras and 3D points).
- The second test (TEST NONRIGID) simulates inter-frames camera poses inaccuracies inherent to SfM algorithm due to noise, outliers, numerical drift, *etc*. A nonrigid perturbation of the global reconstruction is performed by randomly disturbing the inter-frames camera displacements and then regenerating a 3D point cloud.

The amplitude of the perturbations fluctuates between 1% and 10% of the circle radius formed by the camera trajectory. We then apply the four algorithms on those generated reconstructions. The quality of the final reconstructions are measured as the 3D RMS on the camera positions between the ground truth and the estimated poses. We compare the four refinement algorithms in terms of accuracy, convergence speed and frequency. The accuracy is given by the 3D RMS value when the algorithms converged. The convergence frequency is the percentage of trials for which the 3D RMS has decreased. The convergence speed is measured by the 3D error evolution during the LM iterations for a given perturbation magnitude (we use 4% in our experiments). Results are shown on Figure 3. They are mean over 50 random trials.

Convergence frequency. The four algorithms have similar behavior for TEST RIGID and TEST NONRIGID. EG_M&E and BA_M&E have the largest convergence basin

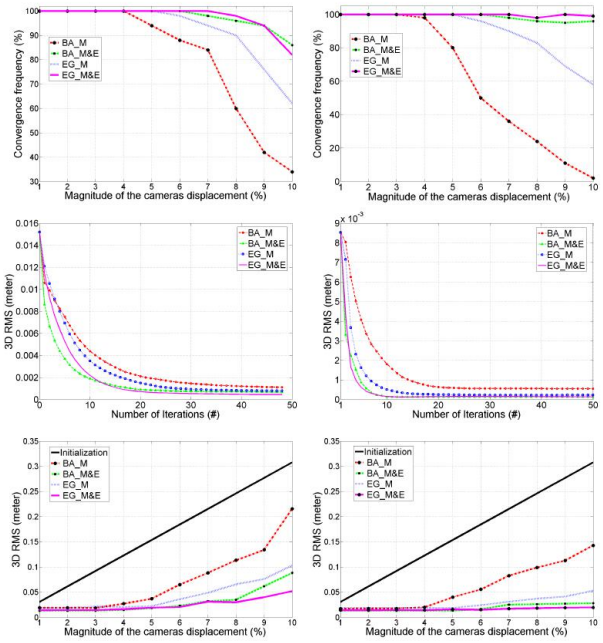

Figure 3. Results obtained with the four algorithms BA_M, BA_M&E, EG_M and EG_M&E for TEST RIGID (Left) and TEST NONRIGID (Right).

where as BA_M has the smallest one. For a displacement magnitude of 10% in TEST NONRIGID, EG_M&E and BA_M&E converge approximatively in all the cases, EG_M converges at 60% and BA_M never converges.

Accuracy. EG_M&E and BA_M&E are the most accurate algorithms against both rigid and nonrigid perturbations. They are closely followed by EG_M whereas BA_M has the worst performance. For example, at 8% of rigid perturbation, the resulting 3D RMS of EG_M&E and BA_M&E are under 5cm. It is over 5cm for EG_M and arround 11cm for BA_M.

Convergence speed. BA_M&E converges faster than the other algorithms. For TEST RIGID, after 3 iterations BA_M&E reduces the 3D error by a factor 2.7, EG_M&E by a factor 2 and EG_M and BA_M by a factor 1.5 only.

Overall. BA_M&E and EG_M&E outperform BA_M and EG_M in terms of accuracy, convergence speed and frequency. It proves that including the unknown part of the environment in the nonlinear refinement improves dramatically the results.

5. Application to Localization in Large Environments and 3D Object Tracking

In this section we evaluate on real data the contribution of including the unknown part of the environment in the nonlinear refinement process for localization in large environment and online 3D object tracking. Two real sequences have been acquired by a low-cost IEEE1394 GUPPY camera providing (640×480) images at 30 frames per second.

5.1. Localization in Large Environments

Recent works, *e.g.* [7], have shown that using a SIG model to constraint SfM point cloud yields accurate reconstruction of urban context. The precision of SIG model, *i.e.* vertical planes roughly representing the building fronts, does not currently allow an accurate online correction of SfM algorithms [8]. However they are globally consistent and thus well suited for a post-processing refinement as in [7]. They propose a two step offline process that roughly aligns the reconstruction on the model with a nonrigid ICP and then refines the reconstruction with model-based nonlinear optimization. Note that the unknown part of the environment *i.e.* everything that are not a building is not taken into account in their optimization process. Their nonlinear refinement algorithm is similar to BA_M and EG_M ones with a slightly different cost function.

Those kind of refinement algorithms are obviously not very robust when the known part of the environment is occluded or absent. However it is very common in urban context: there is not always buildings in each street and they may be occluded by buses, trucks, *etc*. We demonstrate below that using also the unknown part of the environment *i.e.* the road, trees, *etc* in the minimization resolve those challenging issues.

A video sequence of 975 images has been acquired along a 500 meter trip in Versailles, France. The initial SfM reconstruction has been roughly aligned on the model with a NonRigid ICP correction as in [7]. It is then refined by the EG_M&E and EG_M algorithms. The resulting reconstructions are then qualitatively evaluated through online relocation.

Offline nonlinear refinement. Figure 4 shows the 3D point cloud and the camera trajectories obtained after the refinement with EG_M and EG_M&E algorithms. The main differences are localized in two parts of the scene which have been surrounded and zoomed. In the first case, cameras at the beginning of the crossroads do not observe any building whereas for the second one, buildings on the left side are occluded by a bus. EG_M algorithm seems to fail in the two critical cases since the camera presents improbable trajectory and the 3D point cloud improbable structure. On the other hand EG_M&E algorithm results in a smooth trajectory especially in the first area where the structure of a wall seems to be well recovered. These results show that using the unknown part of the environment has improved the overall quality of the reconstruction[5]. These two reconstructions have then been used to build two databases that

[5] Similar conclusions have been obtained with the BA_M&E and BA_M algorithms.

encode the descriptors of all the 3D points[6] in vocabulary tree structures.

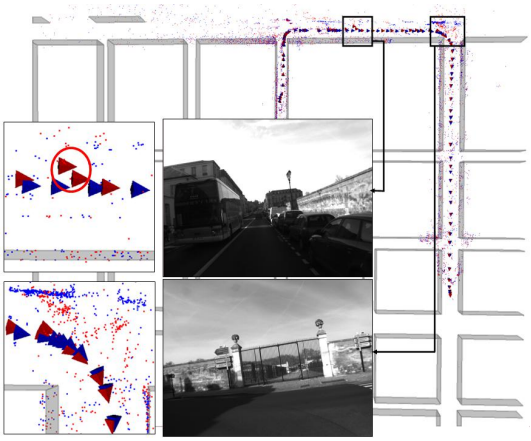

Figure 4. SfM reconstructions of a large outdoor environment. In blue (resp. red) the resulting reconstruction after refinement with the EG_M&E (resp. EG_M) algorithm. Red circles enhance improbable camera positions introduce by the EG_M algorithm.

Online Relocation. Another video sequence of 649 images has been acquired by following approximatively the same trajectory. The environment may have slightly change in the meantime *e.g.* some cars parked have moved. For each image of this sequence a relocation is performed. It associates the descriptors extracted from the query images to those stored in the databases. Then, the camera poses are computed with a RANSAC estimator on the matches. The quality of the relocation is measured as the percentage of inliers when computing the pose through RANSAC. If this percentage is over 60% then the relocation is declared successful. Note that the number of 2D/3D correspondences must also be over a given threshold (10 in our experiments). The number of successful relocation is 539 and 642 for the databases obtained with the EG_M and the EG_M&E algorithms respectively. With the former database, relocation failures appears in the two parts of the sequence described above. This proves that the EG_M&E algorithm yields in a higher quality reconstruction. Moreover, it has also increased the number of the 2D/3D correspondences used to compute the pose: 50 vs 60 in average (computed when relocation in both maps are successful) along the sequence. We can conclude that 3D points associated to the known and unknown parts of the environment are more consistent in the reconstruction refined by the EG_M&E algorithm.

[6]3D points associated to both the known and the unknown parts of the environment are kept in the both maps.

5.2. 3D Object Tracking

The object of interest is a toy car representing the Citroen C4 of Sebastien Loeb in WRC championship. The 3D model used in our experiments is composed by 1600 triangles. It does not include some pieces of the cars such as wheels, airfoil, windows, *etc*. We place the car in a desk context composed by a computer screen, keyboard, books, *etc*. They constitute the unknown part of the environment.

Online tracking. We use the sequential SfM algorithm described in [10]. It is based on a nonlinear refinement algorithm which is applied to a sliding window of triplets of keyframes through local Bundle Adjustment (LBA). At each keyframe, only the poses associated to the three last key-frames and the 3D points they observed are optimized. We modify the LBA to take into account the model constraints. Thus the minimization of Eq. (7) and Eq. (5) has also been implemented in a local Bundle Adjustment framework. We call these refinement algorithms LBA_M&E and LBA_M respectively. Realtime performance is obtained by taking advantage of the sparse block structure of the normal equations associated to both (unknown and known parts of the environment) terms[7]. We compare the three nonlinear refinement algorithms: LBA, LBA_M, LBA_M&E on a challenging sequence. It presents large variation in scale, fast motion, lighting variation, partial and total occlusion of the toy car, *etc*. Coordinate frame registration is performed through matching between the first frame of the sequence and a key-frame registered offline on the model.

Results. Figure 5 presents the results obtained by the three refinement algorithms on this sequence. The coordinate frame registration seems accurate on the first frame (the model is well projected on it) but after turning around the car we observe that this not really the case. The LBA refinement algorithm can not correct this inaccuracy as seen on Figure 5 (Top Left) and thus will conserve it during the whole sequence. LBA_M&E and LBA_M refinement algorithms manage to correct the registration error after few frames: the front and the back of the car are perfectly projected on the images. On the other hand the LBA_M&E refinement algorithm outperforms the LBA_M one when the object of interest is occluded or take a small parts in the images as seen in Figure 5 (Right). It successfully manages to localize, in realtime the toy during the whole sequence. Combing the informations providing by the known and the unknown parts of the environment yields to accurate and robust localization. The LBA_M&E refinement algorithm is then perfectly designed for Augmented Reality applications as illustrated in Figure 6.

[7]EG_M&E and EG_M do not present a such nice property (with more non zero values in the normal equations) and are thus not well suited for real-time performance.

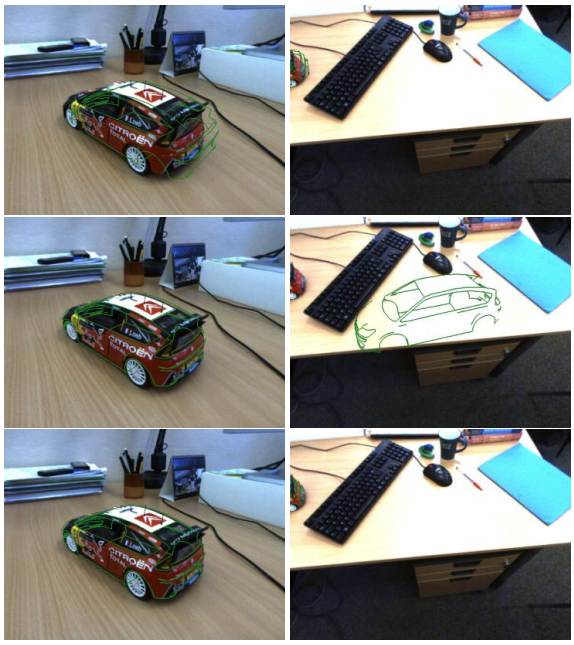

Figure 5. 3D object tracking with Local Bundle Adjustment algorithms. Top, Middle, Bottom: results obtained with the LBA, LBA_M, LBA_M&E refinement algorithms respectively.

Figure 6. LBA_M&E refinement algorithm and Augmented Reality applications.

6. Conclusion

We have presented nonlinear refinement algorithms of an initial SfM reconstruction in a partially known environment. Two compound cost functions that include both informations provided by the known and the unknown parts of the environment have been proposed. Their optimal combination as well as the optimization process have also carefully been studied. Experimental results on both synthetic and real data demonstrate that using the unknown part of the environment in the minimization yields refinement algorithms that outperform those using only the model constraints in terms of accuracy, robustness and convergence basin. We successfully apply our framework to online 3D object tracking and offline outdoor localization for Augmented Reality purposes. Further work will investigate classification improvements between the unknown and known parts of the environment. This is currently done roughly by ray tracing. We expect that image segmentation or keypoint clustering techniques will help for a more accurate classification.

References

[1] A. Bartoli and P. Sturm. Constrained structure and motion from multiple uncalibrated views of a piecewise planar scene. *Int. J. Comput. Vision*, 52:45–64, April 2003. 2

[2] G. Bleser, H. Wuest, and D. Stricker. Online camera pose estimation in partially known and dynamic scenes. In *ISMAR*, 2006. 1

[3] M. Farenzena, A. Bartoli, and Y. Mezouar. Efficient camera smoothing in sequential structure-from-motion using approximate cross-validationt. In *ECCV*, 2008. 4

[4] R. I. Hartley and A. Zisserman. *Multiple View Geometry in Computer Vision*. Cambridge University Press, 2000. 2

[5] G. Klein and D. Murray. Parallel tracking and mapping for small AR workspaces. In *ISMAR*, 2007. 1, 2

[6] V. Lepetit and P. Fua. Monocular model-based 3d tracking of rigid objects: A survey. In *FTCGV*, 2005. 1

[7] P. Lothe, S. Bourgeois, F. Dekeyser, E. Royer, and M. Dhome. Towards geographical referencing of monocular slam reconstruction using 3d city models: Application to real-time accurate vision-based localization. In *CVPR*, 2009. 1, 6

[8] P. Lothe, S. Bourgeois, E. Royer, M. Dhome, and S. Naudet-Collette. Real-time vehicle global localisation with a single camera in dense urban areas: Exploitation of coarse 3d city models. In *CVPR*, 2010. 6

[9] D. Marquardt. An algorithm for least-squares estimation of non linear parameters. *J. Soc. Industr. Appl. Math.*, 11(1):431–444, 1963. 4

[10] E. Mouragnon, M. Lhuillier, M. Dhome, F. Dekeyser, and P. Sayd. Real time localization and 3d reconstruction. In *CVPR*, 2006. 1, 2, 5, 7

[11] D. Nister, O. Naroditsky, and J. Bergen. Visual odometry. In *CVPR*, 2004. 1

[12] E. Royer, M. Lhuillier, M. Dhome, and T. Chateau. Localization in urban environments: Monocular vision compared to a differential gps sensor. In *CVPR*, 2005. 1

[13] G. Simon and M.-O. Berger. Pose estimation for planar structures. *IEEE Computer Graphics and Applications*, 22(6):46–53, 2002. 3

[14] N. Snavely, S. M. Seitz, and R. Szeliski. Modeling the world from internet photo collections. *IJCV*, 80(2):189–210, 2008. 1, 2

[15] R. Szeliski and P. H. S. Torr. Geometrically constrained structure from motion: Points on planes. In *European Workshop on 3D Structure from Multiple Images of Large-Scale Environments*, pages 171–186, 1998. 2

[16] B. Triggs, P. F. McLauchlan, R. I. Hartley, and A. W. Fitzgibbon. Bundle adjustment - a modern synthesis. In *ICCVW: International Workshop on Vision Algorithms Theory and Practice*, 2000. 2