2013 IEEE International Conference on Computer Vision

Elastic Fragments for Dense Scene Reconstruction

Qian-Yi Zhou[1] Stephen Miller[1] Vladlen Koltun[1,2]

[1]Stanford University [2]Adobe Research

Abstract

We present an approach to reconstruction of detailed scene geometry from range video. Range data produced by commodity handheld cameras suffers from high-frequency errors and low-frequency distortion. Our approach deals with both sources of error by reconstructing locally smooth scene fragments and letting these fragments deform in order to align to each other. We develop a volumetric registration formulation that leverages the smoothness of the deformation to make optimization practical for large scenes. Experimental results demonstrate that our approach substantially increases the fidelity of complex scene geometry reconstructed with commodity handheld cameras.

1. Introduction

Enabling the reconstruction of detailed surface geometry from image data is one of the central goals of computer vision. Substantial progress on dense scene reconstruction from photographs and video sequences has been made, despite the ambiguity of photometric cues [20, 26, 21, 6, 15, 17]. When direct information on the surface geometry of the scene is given in the form of range data, we can expect to do even better. The recent commercialization of consumer-grade range cameras promises to enable almost anyone to reliably create detailed three-dimensional models of their environments [27, 16, 7, 35].

Obtaining a detailed three-dimensional model of an object or an environment from range images is difficult in part due to the high-frequency noise and quantization artifacts in the data [11, 27]. This difficulty can be addressed to a significant extent by integrating a large number of range images. Notably, Newcombe et al. [16], building on work on range image integration [2], real-time range scanning [23], and monocular SLAM [3, 4, 12] showed that registering each input image to a growing volumetric model can average out high-frequency error and produce smooth reconstructions of objects and small scenes. These ideas have subsequently been extended to larger environments [34, 35].

A related source of difficulty is the substantial low-frequency distortion present in range images produced by consumer-grade sensors [27, 11, 8]. Even with careful calibration, non-trivial distortion remains. This may not lead to noticeable artifacts if the scanned objects are relatively small or if the scanned surfaces do not contain fine-scale details. However, for sufficiently large and complex scenes this distortion leads to clearly visible artifacts in the reconstructed geometry (Figure 2).

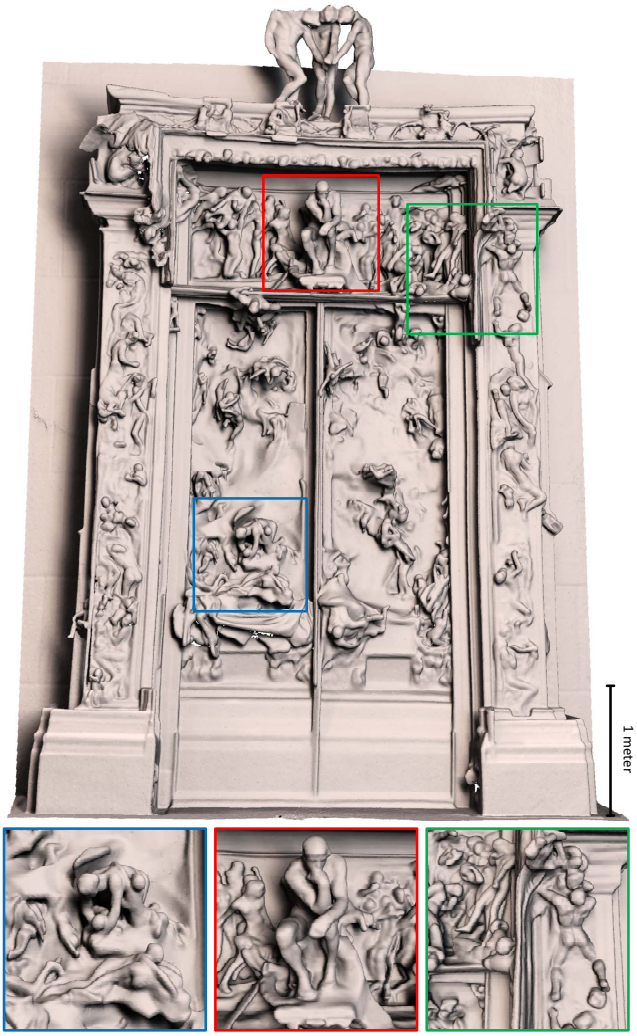
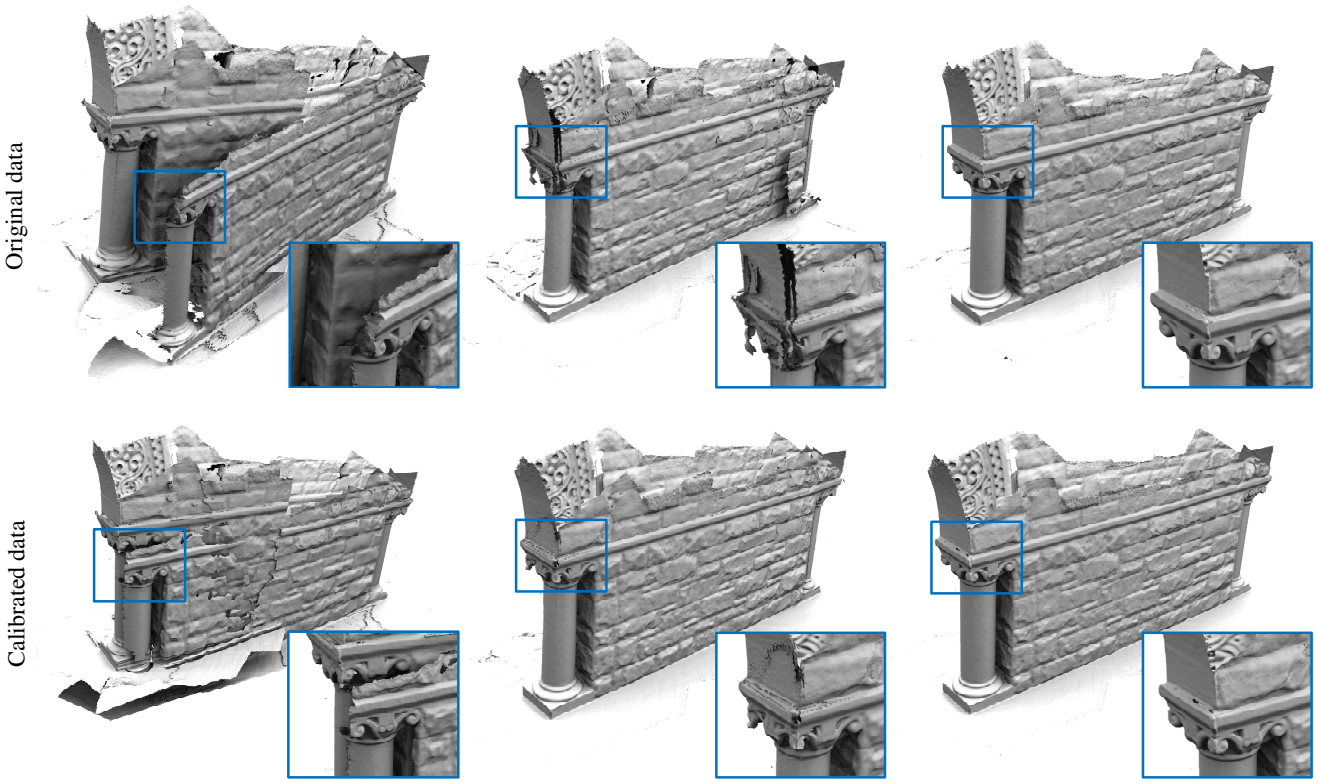


Figure 1. Rodin's "The Gates of Hell," reconstructed from range data acquired with a handheld consumer-grade camera. The sculpture is 4 meters wide and 6 meters high.

1550-5499/13 $31.00 © 2013 IEEE
DOI 10.1109/ICCV.2013.65



(a) Extended KinectFusion (b) Zhou and Koltun (c) Our approach

Figure 2. Reconstructions produced by different approaches on the stone wall sequence from Zhou and Koltun [35]. The top row shows results with original data from the sensor, the bottom row shows results with data that was processed by the calibration approach of Teichman et al. [31], which reduces low-frequency distortion. (a) Extended KinectFusion [22] is unable to produce a globally consistent reconstruction due to drift. (b) The approach of Zhou and Koltun [35] is restricted to rigid alignment and is unable to correct the inconsistencies in trajectory fragments acquired at different times and from different points of view. (c) Our approach uses elastic registration to align corresponding areas on different fragments, producing clean results on both original and calibrated data; the results are virtually identical in the two conditions, indicating that our approach can successfully deal with substantial low-frequency distortion in the input.

Current techniques for dense scene reconstruction from consumer-grade range video cast the problem in terms of trajectory estimation [16, 7, 34, 35]. The implicit assumption is that once a sufficiently accurate estimate for the camera trajectory is obtained, the range images can be integrated to yield a clean model of the scene's geometry. The difficulty is that for sufficiently complex scenes and camera trajectories there may not be any estimate for the trajectory that yields an artifact-free reconstruction with rigidly aligned images, due to the low-frequency distortion in the input. Rigidly aligning the images along a camera path is not always sufficient to resolve the inconsistencies produced by distortions in the sensor.

In this work, we introduce a scene reconstruction approach that is based on non-rigid alignment. Our guiding observation is that we can reliably obtain geometry that is locally accurate. Specifically, we partition the input stream into small fragments of k frames each. Frame-to-model registration [16] is used to reconstruct the surfaces imaged in each fragment, integrating out high-frequency error. Since the low-frequency distortion introduced by the sensor is intrinsically stationary and since the fragments are temporally brief, each fragment is internally consistent. The problem is that fragments that were acquired from substantially different points of view are in general not mutually consistent. Our approach allows the fragments to subtly bend to resolve these extrinsic inconsistencies. This is done by optimizing a global objective that maximizes alignment between overlapping fragments while minimizing elastic strain energy to protect local detail.

Non-rigid registration has a long history in medical imaging and computer vision, resulting in sophisticated techniques for aligning two-dimensional contours and three-dimensional shapes [19, 9, 33, 14, 10]. These techniques primarily aim to align two or more reasonably complete representatives from an object class. Our work aims to reconstruct spatially extended scenes from a large number of range images, each of which covers only a small part of the scene. Real-world scenes can have detailed geometric features at multiple scales. Our approach was thus designed

to preserve surface detail while operating on a scale that has rarely been addressed with non-rigid registration techniques.

The closest work to ours is due to Brown and Rusinkiewicz, who used non-rigid alignment to produce precise object models from 3D scans [1]. We adopt the basic idea of employing non-rigid deformation to preserve surface detail, but develop a different formulation that is more appropriate to our setting. Specifically, the approach of Brown and Rusinkiewicz is based on detecting and aligning keypoints, and propagating this sparse alignment using thin-plate splines. This approach can be problematic because keypoint-based correspondences are imperfect in practice and the spline interpolation is insensitive to surface detail outside the keypoints. We formulate an optimization objective that integrates alignment and regularization constraints that densely cover all surfaces in the scene. Since our input is a temporally dense stream of range data, we can establish correspondences directly on dense geometry without singling out keypoints. This enables the formulation of a regularization objective that reliably preserves surface detail throughout the scene.

Figures 1 and 2 illustrate the benefits of elastic registration. Our approach produces clean reconstructions of large scenes. Since the high-frequency noise is integrated out by individual fragments and the low-frequency distortion is resolved when the fragments are registered to each other, detailed surface geometry is cleanly reconstructed throughout the scene. We demonstrate the effectiveness of the presented approach on a variety of real-world scenes and complex synthetic models.

2. Overview

Fragment construction. We exploit the fact that while online reconstruction methods are unstable over long ranges, they are quite accurate in the local regime. Given an RGB-D scan as input, we partition it into k-frame segments (we use $k = 50$ or $k = 100$) and use the frame-to-model registration and integration pipeline developed by Newcombe et al. [16] to reconstruct a locally precise surface fragment from each such trajectory segment. Each fragment is a triangular mesh with the vertex set $\mathbf{P}_i = \{\mathbf{p}\}$ and the edge set $\mathcal{G}_i \subset \mathbf{P}_i^2$.

Initial alignment. The purpose of initial alignment is to establish dense correspondences between fragments that cover overlapping parts of the scene. To initialize this process, we assume that a rough initial alignment between the fragments in an extrinsic coordinate frame ("scene frame") can be obtained. While prior work relied on manual initial alignment [1], we found that an off-the-shelf SLAM system [5] was sufficient for our purposes. Given the rough localization, we identify pairs of overlapping fragments. To this end, we test every pair of fragments and attempt to align it using ICP starting with the relative pose provided by the rough initialization. If ICP converges with stable correspondences over a sufficiently large area (more than 20% of one of the fragments), we retain the correspondences. Consider such a pair of fragments $(\mathbf{P}_i, \mathbf{P}_j)$. The set of correspondences obtained by ICP that fall below a reasonable global threshold (3cm in all our experiments) are denoted by $\mathcal{K}_{i,j}$. These correspondence sets, established over many pairs of overlapping fragments, are used in the next stage to define the alignment objective.

Elastic registration. Given the correspondences extracted in the preceding stage, we define an optimization objective that combines an alignment term and a regularization term. The alignment term minimizes the distances between corresponding points on different fragments. The regularization term preserves the shape of each fragment by minimizing the elastic strain energy produced by the deformation. A natural formulation of this objective is described in Section 3.1. Unfortunately, this formulation is computationally infeasible for the problems we are dealing with. It also deals poorly with fragments that have multiple connected components, which are commonly encountered in complex scenes. We therefore develop an alternative formulation, described in Section 3.2, that resolves these difficulties. The objective is optimized using an iterative least squares scheme, described in Section 3.3.

Integration. Volumetric integration [2] is used to merge the fragments and to obtain the complete scene model.

3. Elastic Registration

We begin in Section 3.1 by providing a natural point-based formulation of the problem. This is used to introduce the basic structure of optimization objective. After motivating the objective and clarifying the deficiencies of the initial approach, we develop a volumetric formulation that addresses these issues in Section 3.2. The optimization procedure is described in Section 3.3.

3.1. Point-based Registration

Our input is a set of fragments, each parameterized in its own coordinate system. Our objective is to find a mapping $\mathbf{T}$ that maps each point set $\mathbf{P}_i$ to an isomorphic point set $\mathbf{P}'_i$, such that all sets $\mathbf{P}'_i$ are parameterized in a common coordinate frame and are aligned to form a global model of the scanned scene. Let $\mathbf{P} = \bigcup \mathbf{P}_i$ be the set of all input points and let $\mathbf{P}' = \bigcup \mathbf{P}'_i$ be the corresponding output set. The desired mapping $\mathbf{T}$ should minimize the distance between corresponding point pairs $\mathcal{K}_{i,j}$ for all i, j while preserving the detailed geometry of each fragment. We compute $\mathbf{T}$ by

minimizing an energy function of the form

$$E(\mathbf{T}) = E_a(\mathbf{T}) + E_r(\mathbf{T}), \tag{1}$$

where E_a is the alignment term and E_r is the elastic regularization term.

The alignment term $E_a(\mathbf{T})$ measures the alignment of all corresponding pairs. We use the point-to-plane distance, which has well-known benefits for surface registration [24]:

$$E_a(\mathbf{T}) = \sum_{i,j} \sum_{(\mathbf{p},\mathbf{q})\in\mathcal{K}_{i,j}} \|(\mathbf{p}' - \mathbf{q}') \cdot \mathbf{N}'_{\mathbf{p}}\|^2. \tag{2}$$

In our notation, $\mathbf{p}' = \mathbf{T}(\mathbf{p})$ and $\mathbf{N}'_{\mathbf{p}}$ is the normal of $\mathbf{P}'_i$ at $\mathbf{p}'$. $\mathcal{K}_{i,j} = \emptyset$ if no correspondences between $\mathbf{P}_i$ and $\mathbf{P}_j$ were established.

The regularizer $E_r(\mathbf{T})$ measures the elastic strain energy for all fragments [32]. In principle, we want to measure the change in the first two fundamental forms of each surface due to the mapping $\mathbf{T}$:

$$\sum_i \int_{\Omega_i} k_s \|\mathbf{I}'_i - \mathbf{I}_i\|^2 + k_b \|\mathbf{II}'_i - \mathbf{II}_i\|^2 du dv, \tag{3}$$

where $\mathbf{I}_i, \mathbf{I}'_i$ and $\mathbf{II}_i, \mathbf{II}'_i$ are the first and second fundamental forms of $\mathbf{P}_i, \mathbf{P}'_i$, respectively, and k_s and k_b are stiffness parameters. For mild low-frequency deformations, which are the kinds of deformations induced by optical distortion, (3) can be approximated as follows:

$$E_r(\mathbf{T}) = \sum_i \sum_{\mathbf{p}\in\mathbf{P}_i} \sum_{\mathbf{q}\in\mathcal{N}(\mathbf{p})} \|(\mathbf{q}' - \mathbf{p}') - \mathbf{R}'_{\mathbf{p}}(\mathbf{q} - \mathbf{p})\|^2, \tag{4}$$

where $\mathcal{N}(\mathbf{p})$ is the set of neighbors of $\mathbf{p}$ in $\mathcal{G}_i$, and $\mathbf{R}'_{\mathbf{p}}$ is a rotation transform that maps the local tangent frame of $\mathbf{p}$ to the local tangent frame of $\mathbf{p}'$ [28, 30]. $\mathbf{R}'_{\mathbf{p}}$ thus represents the local rotational effect of $\mathbf{T}$ at $\mathbf{p}$. The intuition behind this formulation is that $\mathbf{R}'_{\mathbf{p}}$ corrects for the rotation induced by $\mathbf{T}$, so the linear least-squares term $\|(\mathbf{q}' - \mathbf{p}') - \mathbf{R}'_{\mathbf{p}}(\mathbf{q} - \mathbf{p})\|^2$ penalizes distortion induced by $\mathbf{T}$ over the edge $(\mathbf{p}, \mathbf{q})$. Since $\mathbf{R}'_{\mathbf{p}}$ is used to rotationally align $\mathbf{q} - \mathbf{p}$ with $\mathbf{q}' - \mathbf{p}'$, this linear term conveniently penalizes both stretching and bending of $\mathbf{P}_i$ at $(\mathbf{p}, \mathbf{q})$.

If the transformed normal $\mathbf{N}'_{\mathbf{p}}$ and the tangent frame rotation $\mathbf{R}'_{\mathbf{p}}$ are known, both $E_a(\mathbf{T})$ and $E_r(\mathbf{T})$ are linear least-squares objectives that can in principle be solved efficiently. Of course, neither $\mathbf{N}'_{\mathbf{p}}$ nor $\mathbf{R}'_{\mathbf{p}}$ are known in advance because both depend on the transform $\mathbf{T}$, which is being optimized. However, this suggests an iterative optimization scheme. In each step, we fix $\mathbf{N}'_{\mathbf{p}}$ and $\mathbf{R}'_{\mathbf{p}}$ and solve for $\mathbf{T}$ (specifically, for the point set $\mathbf{P}'$ that minimizes (1)). We then compute an updated estimate for the local normal and tangent frame at each point $\mathbf{p}' \in \mathbf{P}'$ and repeat.

Since each step involves simply solving a linear least squares problem, we can expect this procedure to be efficient. While in principle it is, the scale of our scenes makes it impractical. For example, the scene shown in Figure 1 contains 370 fragments with a total of 66.5 million points, yielding a linear system with 199 million variables and 7.8 trillion non-zero entries in the matrix. Furthermore, the point-based formulation does not control for distortion induced by changes in the relative pose of disconnected surfaces within fragments. In Section 3.2, we reformulate the registration objective to address these issues.

3.2. Volumetric Registration

The guiding observation behind the reformulation is that the unknown transform $\mathbf{T}$ is assumed to be smooth (low-frequency) over the domain of each fragment. This function can thus be evaluated at a small number of samples and reconstructed by interpolation. We thus embed each fragment $\mathbf{P}_i$ in a coarse control lattice $\mathbf{V}_i$. The mapping $\mathbf{T}$ is defined for $\mathbf{V} = \bigcup \mathbf{V}_i$ and is applied to $\mathbf{P}$ by interpolation. Specifically, let the set of vertices in $\mathbf{V}_i$ be $\{\mathbf{v}_{i,l}\}$. Each point $\mathbf{p} \in \mathbf{P}_i$ can be represented as a linear combination of vertices from $\mathbf{V}_i$:

$$\mathbf{p} = \sum_l c_l(\mathbf{p}) \mathbf{v}_{i,l},$$

where $\{c_l(\mathbf{p})\}$ are interpolation coefficients precomputed from a set of basis functions centered at the corresponding control points [25]. These coefficients remain constant during the optimization. The point $\mathbf{p}' = \mathbf{T}(\mathbf{p})$ is reconstructed from the transformed control points $\mathbf{v}' = \mathbf{T}(\mathbf{v})$ by the same interpolation:

$$\mathbf{p}' = \sum_l c_l(\mathbf{p}) \mathbf{v}'_{i,l}. \tag{5}$$

The optimization objective (1) is redefined for $\mathbf{V}$. To this end, we need to reformulate each objective term on $\mathbf{V}'$ instead of $\mathbf{P}'$. This is a simple application of Equation 5. Let's consider the alignment term first. To formulate $E_a(\mathbf{V}')$, we simply substitute (5) into (2). This yields a linear least-squares problem on $\mathbf{V}'$.

To reformulate the regularization term $E_r(\mathbf{T})$ and also address the potential issues due to discontinuities within fragments, we define this term directly on the control lattice:

$$E_r(\mathbf{V}') = \sum_i \sum_{\mathbf{v}\in\mathbf{V}_i} \sum_{\mathbf{u}\in\mathcal{N}_{\mathbf{v}}} \|(\mathbf{u}' - \mathbf{v}') - \mathbf{R}'_{\mathbf{v}}(\mathbf{u} - \mathbf{v})\|^2. \tag{6}$$

3.3. Optimization

If $\{\mathbf{N}'_{\mathbf{p}}\}$ and $\{\mathbf{R}'_{\mathbf{v}}\}$ are fixed, the overall objective $E(\mathbf{V}') = E_a(\mathbf{V}') + E_r(\mathbf{V}')$ is a quadratic function of $\mathbf{V}'$:

$$E(\mathbf{V}') = \|\mathbf{A}\mathbf{V}' - \mathbf{b}\|^2, \tag{7}$$

which can be minimized by solving a linear system:

$$(\mathbf{A}^\top \mathbf{A})\mathbf{V}' = \mathbf{A}^\top \mathbf{b}. \tag{8}$$

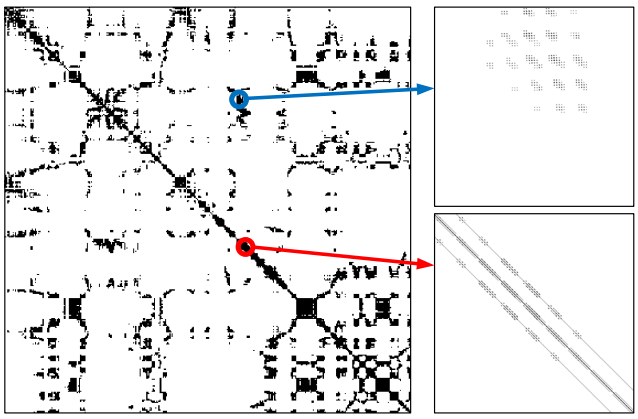


Figure 3. The sparse structure of matrix $\mathbf{A}^\top\mathbf{A}$. The non-zero blocks are shown as black dots in the left image. The internal sparsity of different blocks is visualized on the right. This matrix is for the sculpture in Figure 1.

The matrix $\mathbf{A}^\top\mathbf{A}$ is a block matrix with blocks of size $m \times m$, where $m/3$ is the number of vertices in each control lattice. The regularization term E_r only leads to non-zero entries in the main diagonal blocks of $\mathbf{A}^\top\mathbf{A}$. The alignment term E_a leads to non-zero values in the main diagonal blocks and in blocks that correspond to overlapping fragment pairs for which correspondences were established during the initial alignment stage. (That is, pairs i, j for which $\mathcal{K}_{i,j} \neq \emptyset$.) For large scenes, each fragment will overlap with a constant number of fragments on average and the matrix $\mathbf{A}^\top\mathbf{A}$ will be sparse, as illustrated in Figure 3. In addition, E_a associates control points that jointly participate in determining the position of a point $\mathbf{p}'$, as specified in (5). Given that the basis function anchored at each control point has only local support, each non-zero block in the matrix is internally sparse, as illustrated in Figure 3. To reduce the scale of the problem further we use trilinear interpolation, which associates a lattice vertex only with points in its eight neighboring cells.

The optimization is initialized using the rough rigid alignment computed in the initial alignment stage (Section 2). We estimate the normals $\{\mathbf{N}'_\mathbf{p}\}$ and tangent frame rotations $\{\mathbf{R}'_\mathbf{v}\}$ for this initial configuration. We then proceed with a variant of the iterative optimization scheme outlined in Section 3.1. In each step, we solve the linear system (8). Since $\mathbf{A}^\top\mathbf{A}$ is sparse and symmetric positive definite, we use sparse Cholesky factorization.

Since updating $\{\mathbf{R}'_\mathbf{v}\}$ only affects the right hand side of (8), the linear system can be solved again after such an update using the same factorization of the left hand size. On the other hand, an updated estimate for $\{\mathbf{N}'_\mathbf{p}\}$ changes $\mathbf{A}^\top\mathbf{A}$ and calls for recomputing the Cholesky factorization. We thus update the estimated normals only once per 10 iterations. We perform 50 iterations in total.

4. Experiments

Figures 1, 2, and 4 show the results of our approach on three real-world scenes. We use an Asus Xtion Pro Live camera, which streams VGA-resolution depth and color images at 30 fps. We try to scan as much surface detail as possible in order to evaluate the quality of the reconstruction. A typical scan lasts for 2 to 20 minutes, along a complicated camera trajectory with numerous loop closures. During scanning, the operator could see the color and depth images captured by the sensor in real time, but no preview of the reconstruction was shown.

To evaluate our approach on independently acquired data, we compare our results to three alternatives on the challenging "fr3/long_office_household" scene from the RGB-D SLAM benchmark [29]. The results are shown in Figure 5. Our approach creates a globally consistent scene with high-fidelity local details, while Extended KinectFusion suffers from lack of loop closure and the rigid registration approach of Zhou and Koltun breaks some local regions due to unresolved residual distortion. We also compare to a reconstruction produced by a hypothetical approach that integrates along the motion-captured camera trajectory provided by the benchmark. Despite having access to a motion-captured camera trajectory, this hypothetical approach produces results that are similar to those of Zhou and Koltun. This can be attributed to two potential causes: the approach is limited to rigid alignment and does not resolve the low-frequency distortion in the data, and the sensor noise of the motion capture system.

To further identify the error source and to make quantitative evaluations, we synthesize range video sequences using synthetic 3D models and use these models as ground truth geometry to evaluate the reconstruction quality. To synthesize these sequences, we navigate a virtual camera around each synthetic model and produce perfect range images at full frame rate using a standard rendering pipeline. These images are then combined with two error models to simulate the data produced by real-world range cameras. The two error models we use aim to simulate the factory-calibrated data produced by PrimeSense sensors and idealized data with no low-frequency distortion. To produce the idealized data, we process the perfect synthetic depth images using the quantization model described by Konolige and Mihelich [13] and introduce sensor noise following the model of Nguyen et al. [18]. To produce the simulated factory-calibrated data, we add a model of low-frequency distortion estimated on a real PrimeSense sensor using the calibration approach of Teichman et al. [31].

The results of the synthetic evaluation are shown in Figure 6. The results are obtained by computing the point-to-plane distance from points in the reconstructed model to the ground truth shape, after initial alignment by standard rigid registration. We compare our approach to three alternatives:

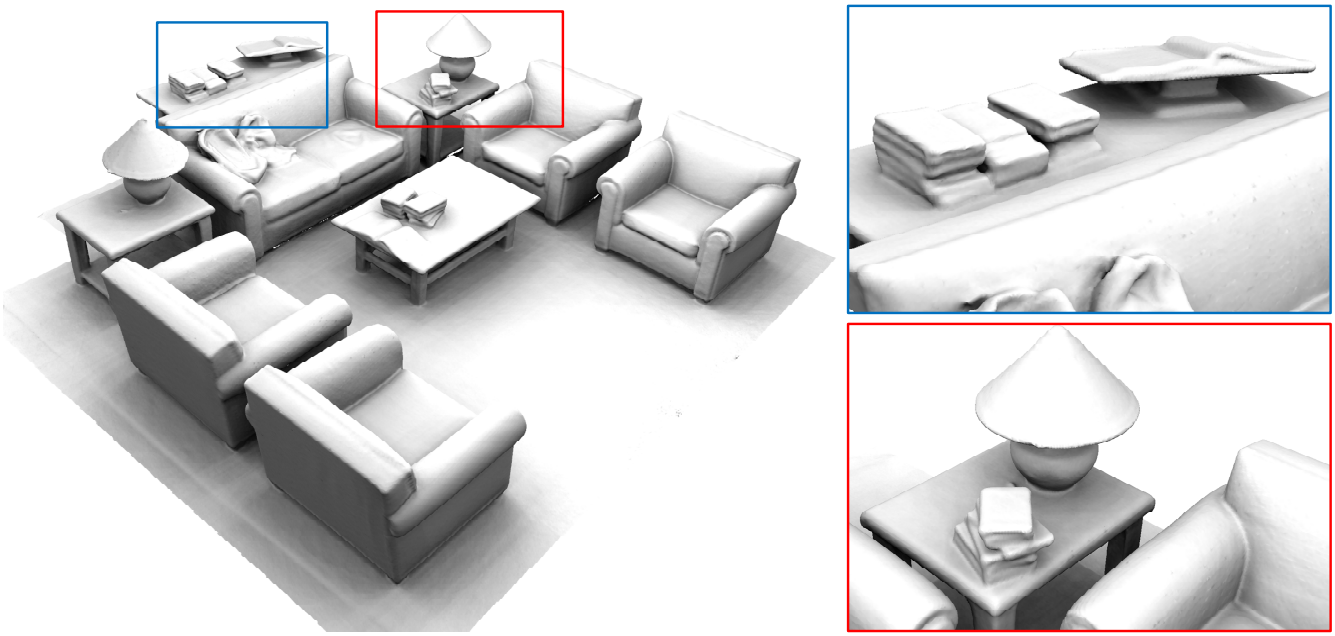


Figure 4. Reconstruction of a sitting area.

Extended KinectFusion [22], Zhou and Koltun [35], and integration of the simulated depth images along the ground truth trajectory. The last alternative is of course inaccessible in practice since the precise camera trajectory is not known, but it is instructive as an experimental condition that isolates reconstruction errors caused by distortion in the input images from reconstruction errors caused by drift in the estimated camera pose.

The results indicate that our approach outperforms both prior approaches (Extended KinectFusion and Zhou and Koltun) with both types of data. For idealized data with no low-frequency distortion, the idealized approach that uses the ground-truth trajectory performs extremely well and outperforms our approach. For simulated factory-calibrated data, our approach sometimes outperforms the idealized approach. This is because the idealized approach is limited to rigid alignment. Although it benefits from perfect camera localization, the real-world distortion in the data causes inconsistencies between input images that are too large to be eliminated by volumetric integration. Our approach uses nonrigid alignment to resolve these inconsistencies.

5. Conclusion

We presented an approach for dense scene reconstruction from range video produced by consumer-grade cameras. Our approach partitions the video sequence into segments, uses frame-to-model integration to reconstruct locally precise scene fragments from each segment, establishes dense correspondences between overlapping fragments, and optimizes a global objective that aligns the fragments. The optimization can subtly deform the fragments, thus correcting inconsistencies caused by low-frequency distortion in the input images.

The approach relies on a number of components that can fail, causing the approach to fail. Current consumer-grade range cameras can fail in the presence of translucent surfaces or direct sunlight. Frame-to-model integration can fail due to jerky camera movement. The SLAM system that we rely on can fail without distinctive and stable visual features. Improving the robustness of these components is a valuable research direction.

References

[1] B. J. Brown and S. Rusinkiewicz. Global non-rigid alignment of 3-D scans. *ACM Trans.Graph.*, 26(3), 2007. 3

[2] B. Curless and M. Levoy. A volumetric method for building complex models from range images. In *SIGGRAPH*, 1996. 1, 3

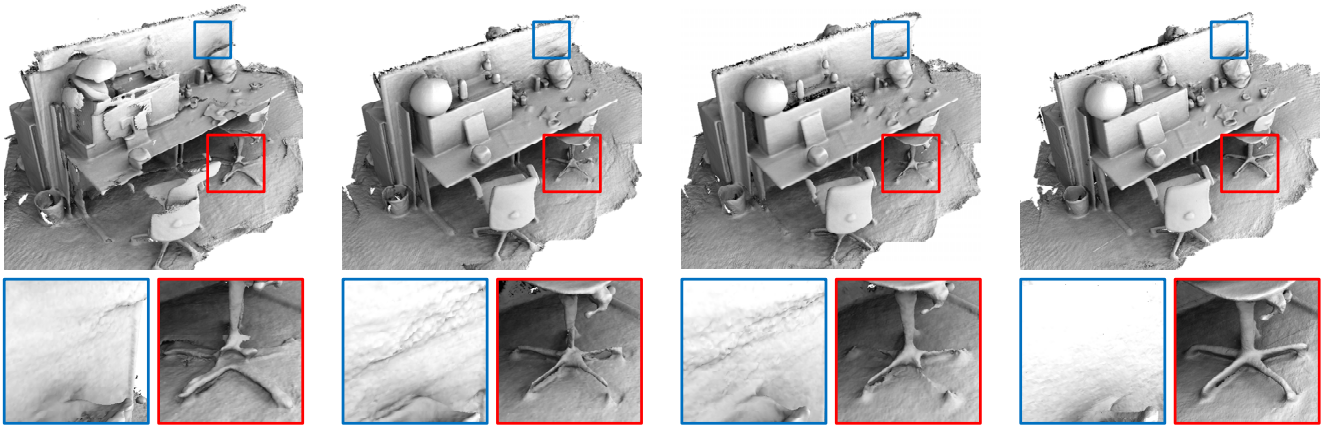
[3] A. J. Davison. Real-time simultaneous localisation and mapping with a single camera. In *ICCV*, 2003. 1

[4] A. J. Davison, I. D. Reid, N. Molton, and O. Stasse. MonoSLAM: Real-time single camera SLAM. *PAMI*, 29(6), 2007. 1

[5] F. Endres, J. Hess, N. Engelhard, J. Sturm, D. Cremers, and W. Burgard. An evaluation of the RGB-D SLAM system. In *ICRA*, 2012. 3

[6] Y. Furukawa, B. Curless, S. M. Seitz, and R. Szeliski. Towards internet-scale multi-view stereo. In *CVPR*, 2010. 1

[7] 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. *International Journal of Robotics Research*, 31(5), 2012. 1, 2



(a) Extended KinectFusion (b) Zhou and Koltun (c) Mocap trajectory (d) Our approach

Figure 5. Evaluation on a benchmark scene [29]: (a) Extended KinectFusion [22], (b) Zhou and Koltun [35], (c) volumetric integration along the motion-captured camera trajectory, and (d) our approach. Our approach is the only one that preserves high-frequency features such as the chair leg (red closeup) without introducing noisy artifacts on the flat panel (blue closeup).

[8] D. Herrera C., J. Kannala, and J. Heikkilä. Joint depth and color camera calibration with distortion correction. *PAMI*, 34(10), 2012. 1

[9] X. Huang, N. Paragios, and D. N. Metaxas. Shape registration in implicit spaces using information theory and free form deformations. *PAMI*, 28(8), 2006. 2

[10] B. Jian and B. C. Vemuri. Robust point set registration using gaussian mixture models. *PAMI*, 33(8), 2011. 2

[11] K. Khoshelham and S. O. Elberink. Accuracy and resolution of Kinect depth data for indoor mapping applications. *Sensors*, 12(2), 2012. 1

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

[13] K. Konolige and P. Mihelich. *Technical description of Kinect calibration*. 2012. http://wiki.ros.org/kinect_calibration/technical. 5

[14] A. Myronenko and X. B. Song. Point set registration: Coherent point drift. *PAMI*, 32(12), 2010. 2

[15] R. A. Newcombe and A. J. Davison. Live dense reconstruction with a single moving camera. In *CVPR*, 2010. 1

[16] 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. In *ISMAR*, 2011. 1, 2, 3

[17] R. A. Newcombe, S. Lovegrove, and A. J. Davison. DTAM: Dense tracking and mapping in real-time. In *ICCV*, 2011. 1

[18] C. V. Nguyen, S. Izadi, and D. Lovell. Modeling Kinect sensor noise for improved 3D reconstruction and tracking. In *3DIMPVT*, 2012. 5

[19] N. Paragios, M. Rousson, and V. Ramesh. Non-rigid registration using distance functions. *CVIU*, 89(2-3), 2003. 2

[20] M. Pollefeys, L. J. V. Gool, M. Vergauwen, F. Verbiest, K. Cornelis, J. Tops, and R. Koch. Visual modeling with a hand-held camera. *IJCV*, 59(3), 2004. 1

[21] M. Pollefeys, D. Nistér, J.-M. Frahm, A. Akbarzadeh, P. Mordohai, B. Clipp, C. Engels, D. Gallup, S. J. Kim, P. Merrell, C. Salmi, S. N. Sinha, B. Talton, L. Wang, Q. Yang, H. Stewénius, R. Yang, G. Welch, and H. Towles. Detailed real-time urban 3D reconstruction from video. *IJCV*, 78(2-3), 2008. 1

[22] H. Roth and M. Vona. Moving volume KinectFusion. In *British Machine Vision Conference (BMVC)*, 2012. 2, 6, 7

[23] S. Rusinkiewicz, O. Hall-Holt, and M. Levoy. Real-time 3D model acquisition. *ACM Trans. Graph.*, 21(3), 2002. 1

[24] S. Rusinkiewicz and M. Levoy. Efficient variants of the ICP algorithm. In *3DIM*, 2001. 4

[25] T. W. Sederberg and S. R. Parry. Free-form deformation of solid geometric models. In *SIGGRAPH*, 1986. 4

[26] S. M. Seitz, B. Curless, J. Diebel, D. Scharstein, and R. Szeliski. A comparison and evaluation of multi-view stereo reconstruction algorithms. In *CVPR*, 2006. 1

[27] J. Smisek, M. Jancosek, and T. Pajdla. 3D with Kinect. In *ICCV Workshops*, 2011. 1

[28] O. Sorkine and M. Alexa. As-rigid-as-possible surface modeling. In *Symposium on Geometry Processing*, 2007. 4

[29] J. Sturm, N. Engelhard, F. Endres, W. Burgard, and D. Cremers. A benchmark for the evaluation of RGB-D SLAM systems. In *IROS*, 2012. 5, 7

[30] R. W. Sumner, J. Schmid, and M. Pauly. Embedded deformation for shape manipulation. *ACM Trans.Graph.*, 26(3), 2007. 4

[31] A. Teichman, S. Miller, and S. Thrun. Unsupervised intrinsic calibration of depth sensors via SLAM. In *RSS*, 2013. 2, 5

[32] D. Terzopoulos, J. C. Platt, A. H. Barr, and K. W. Fleischer. Elastically deformable models. In *SIGGRAPH*, 1987. 4

[33] F. Wang, B. C. Vemuri, A. Rangarajan, and S. J. Eisenschenk. Simultaneous nonrigid registration of multiple point sets and atlas construction. *PAMI*, 30(11), 2008. 2

[34] T. Whelan, H. Johannsson, M. Kaess, J. Leonard, and J. McDonald. Robust real-time visual odometry for dense RGB-D mapping. In *ICRA*, 2013. 1, 2

[35] Q.-Y. Zhou and V. Koltun. Dense scene reconstruction with points of interest. *ACM Trans. Graph.*, 32(4), 2013. 1, 2, 6, 7

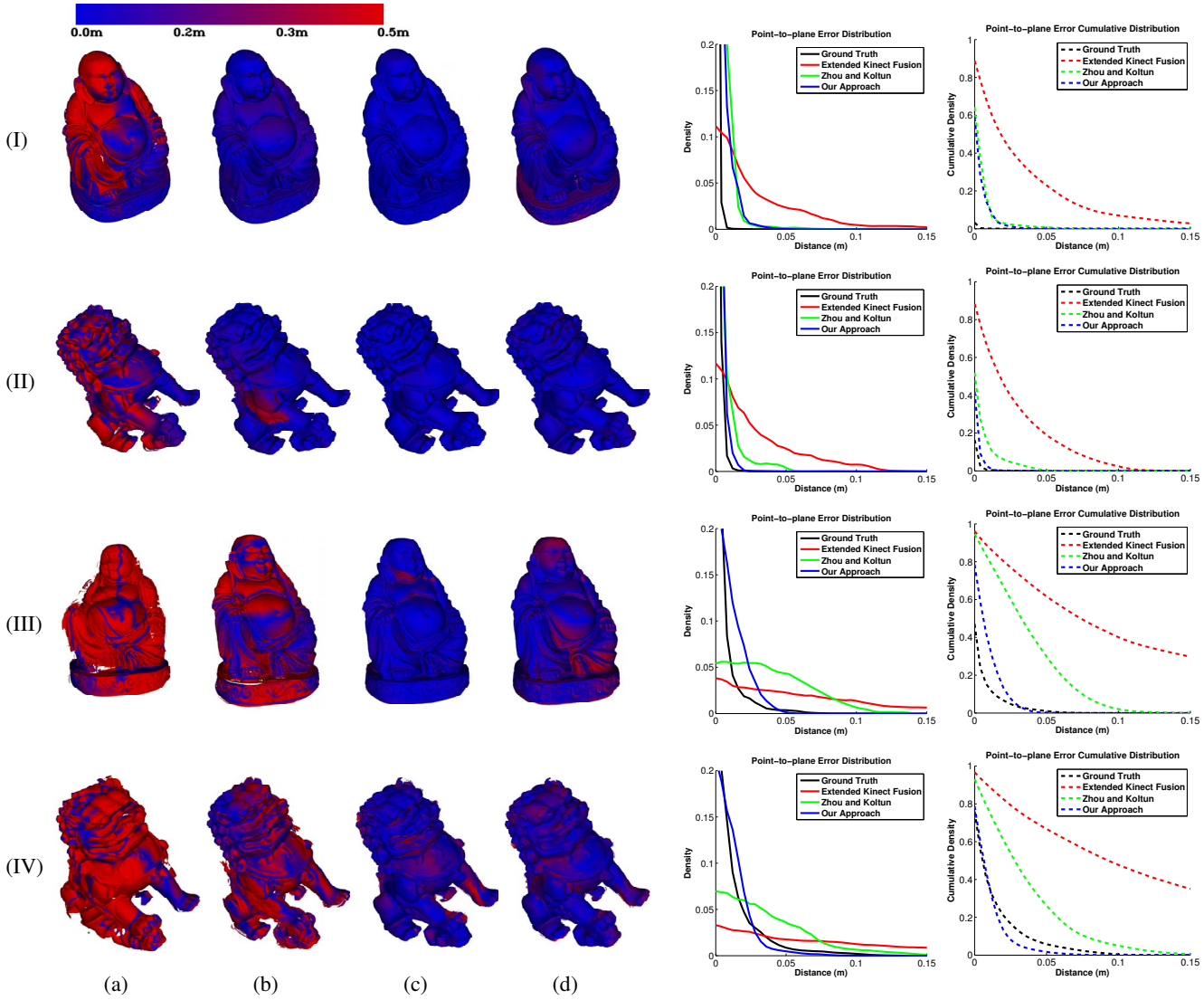


Figure 6. Evaluation with synthetic data. (a) Extended KinectFusion, (b) Zhou and Koltun, (c) volumetric integration along the ground-truth camera trajectory, and (d) our approach. The plots on the right show distributions of point-to-plane error between the reconstructed shapes and the true shape. (I) and (II) use an idealized error model with no low-frequency distortion. (III) and (IV) use the full error model with low-frequency distortion estimated on a real PrimeSense sensor.

Model	Size (L×W×H)	# of fragments	Data collection	Fragment creation	RGB-D SLAM	Initial alignment	Elastic registration	Integration	Total time	Triangle count
Figure 1	4×2×6	370	21m	1h 2m	8h 34m	25h 36m	19h 42m	57m	56h 12m	5,489,745
Figure 2	7×2.5×3.2	54	2m	7m	23m	50m	57m	6m	2h 25m	3,720,310
			2m	7m	23m	57m	59m	6m	2h 34m	3,754,826
Figure 4	5×4.5×1.3	130	7m	24m	2h 12m	6h 27m	7h 30m	23m	17h 3m	3,128,245
Figure 5	4.5×3×1.4	50	-	7m	13m	2h 8m	51m	4m	3h 23m	2,594,934
Figure 6.I	3.5×3.6×5	69	-	10m	-	3h 56m	1h 8m	6m	5h 20m	4,710,726
Figure 6.II	4.6×2.7×5	57	-	8m	-	2h 56m	50m	6m	4h	6,182,782
Figure 6.III	3.5×3.6×5	69	-	10m	-	4h 26m	1h10m	8m	5h 54m	5,115,442
Figure 6.IV	4.6×2.7×5	57	-	8m	-	2h 18m	29m	8m	3h 3m	7,795,335

Table 1. Statistics for the experiments. Length in meters. Running times are measured on a workstation with an Intel i7 3.2GHz CPU, 24GB of RAM, and an NVIDIA GeForce GTX 690 graphics card.