Seminar Report on Volumetric Shading-based Refinement

Meng Liu

Robotics, Cognition, Intelligence- Technische Universität München

Abstract

In this report, the paper Shading-based Refinement on Volumetric Signed Distance Function [5] from Zollhöfer et al. in year 2015 will be discussed in detail. This paper proposes a novel method to efficiently achieve fine-scale 3D reconstruction with commodity sensors. As the reconstruction from low-budget depth sensors often results in over-smoothness, the author suggests using the color images to further elevate the quality of the original fusion. The method jointly optimizes the geometry and albedo on the implicit surface representation, i.e., signed distance functions. A tailored Gauss-Newton solver is presented to speed up the optimization on a voxel hashing structure. This approach has good performances but also lack the ability on large-scale scene or non-Lambertian surface reconstructions.

1 Introduction

Nowadays the commodity RGB-D sensors are ubiquitous, researches like KinectFusion [2] are able to reconstruct surfaces of objects based on these low-budget cameras with some amount of details. However, there is an important limitation of these economical depth sensors, i.e., the low-quality outputs. The proposed method aims to solve this problem by taking the advantage of color images, namely the high resolution. To do so, the method chooses the implicit surface representation, truncated signed distance functions (TSDFs), to store the 3D geometry information. Unlike meshes, one of the explicit surface representation methods, TSDFs are easy to progressively integrate without dealing complex topological connections. Therefore, for computational efficiency TSDFs are preferred. Nevertheless, this structure also leads to over-smoothing because it needs weighted averaging to integrate new geometric information into the existed reconstruction. Shading-based refinement is their main contribution to tackle the problem from initial fusion, and it performs in a hierarchical manner under the voxel hashing scheme. To solve this non-linear optimization problem efficiently, a GPU-based Gauss-Newton solver is crafted. To summary, this report will be presented in the following steps:

- an introduction on the related works is presented in section 2
- a overview of the method is presented in section 3
- some discussion of experiments are in sections 4 and 5
- finally followed by a summary (section 6).

2 Related works

The proposed paper is based on some of previous works [3][4], and meanwhile addresses their main problems. The approach MESHRef [3] operates refinement using color images as well; however, it chose meshes instead of SDFs, which costs lots of computing time to handle the topology. The other method [4] estimates the lighting for every single image which yield inconsistency for different views. Moreover, it refines depth maps independently before fusion, and this pipeline will conversely smooth out the refined details in previous steps.

After two years when the mainly concerned paper was published, a new approach [1] points out some issues that may downgrade its performance. The new method proposes to jointly optimize geometry, albedo, camera poses, camera intrinsics and scene lighting, instead of fixed intrinsics and camera positions after calibration and bundle adjustments used in the main paper. The new approach also suggests implementing a much more flexible spatially-varying Spherical Harmonics to map the lighting model more precisely.

3 Method description

3.1 Pipeline

As shown in the figure 1, the proposed method takes RGB-D images as input, first computes the camera pose for each frame, then fuses the depth frames into TSDFs. Given the fusion, a shading-based refinement is performed in a hierarchical manner, namely coarse-to-fine. The output will be a fine-scale reconstruction. The following part of this section will show details of the method.

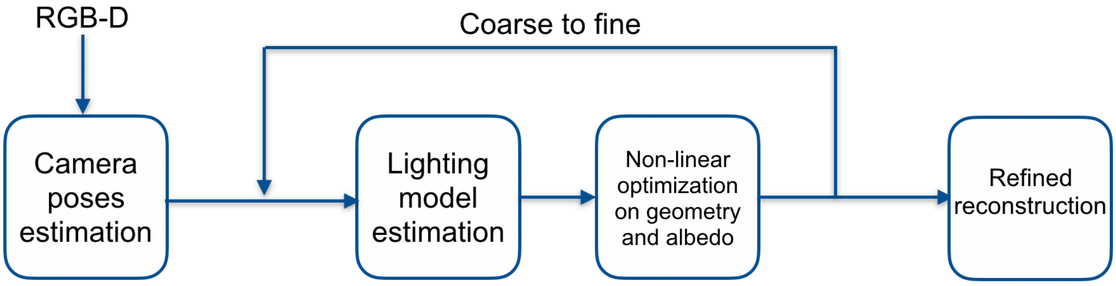
RGB-D
Coarse to fine
Camera poses estimation
Lighting model estimation
Non-linear optimization on geometry and albedo
Refined reconstruction

Figure 1: Overview of the pipeline

1. The very first step of the pipeline is camera poses estimation using bundle adjustment. To achieve fast convergence, the method splits the process into two steps, namely a sparse bundle adjustment step and a dense bundle adjustment step. The sparse bundle adjustment requires detected features as inputs, and the equation 1 has an interpretation that the features seen from different viewpoints should have similar world coordinates. T and $\boldsymbol{p}$ represent camera pose and feature point.

$$E_{sparse}(T) = \sum_{i,j}^{\#frames} \sum_{k}^{\#corresp.} \|T_i \boldsymbol{p}_{ik} - T_j \boldsymbol{p}_{jk}\|_2^2 \tag{1}$$

However, the precision of poses got from this sparse term is relative low, as the feature extracted from SIFT corner detector may have some pixels off, therefore a dense bundle adjustment is needed as shown in eqs. (2) to (4).

$$E_{dense}(T) = w_{color} E_{color}(T) + w_{geometry} E_{geometry}(T) \tag{2}$$

$$E_{color}(T) = \sum_{i,j}^{\#frames} \sum_{k}^{\#pixels} \|I_i(\pi_c(\boldsymbol{p}_{ik})) - I_j(\pi_c(\boldsymbol{p}_{jk}))\|_2^2 \tag{3}$$

Color term represents the photometric consistency, namely the color value of the same 3D point projection in every image pair should be similar,

$$E_{geometry}(T) = \sum_{i,j}^{\#frames} \sum_{k}^{\#pixels} [\boldsymbol{n}_{ik}^T \cdot (\boldsymbol{p}_{ik} - T_i^{-1} T_j \pi_d^{-1}(D_j(\pi_d(T_j^{-1} T_i \boldsymbol{p}_{ik}))))]^2 \tag{4}$$

and the geometry term represents a point-to-plane energy, namely the corresponding 3D coordinate of a pixel should fit into the local geometry when projected to another camera frame.

2. After computed all the camera poses, the approach first fuses all depth maps and initializes the TSDFs, then followed by the lighting model estimation.

$$E_{light}(\boldsymbol{l}) = \sum_{\boldsymbol{v} \in D_0} (B(\boldsymbol{v}) - \boldsymbol{I}(\boldsymbol{v})) \tag{5}$$

$$B(\boldsymbol{v}) = \boldsymbol{a}(\boldsymbol{v}) \sum_{m=1}^{b^2} l_m H_m(\boldsymbol{n}(\boldsymbol{v})) \tag{6}$$

$B(\boldsymbol{v})$ represents the shading on the iso-surface, H_m are the 3 band spherical harmonics basis function, and l_m are the corresponding coefficients. The energy term eq. (5) minimizes the difference between computed shading and intensity at voxel $\boldsymbol{v}$.

3. The final step is non-linear optimization on geometry and albedo of every voxel near the ios-surface. The objective function eq. (7) contains four terms, a fitting term and three regularizers eqs. (8) to (11).

$$E_{refine}(\tilde{\boldsymbol{D}}, \boldsymbol{a}) = \sum_{\boldsymbol{v} \text{ s.t.} |\tilde{\boldsymbol{D}}(\boldsymbol{v})| < t_{shell}} w_g E_g(\boldsymbol{v}) + w_r E_r(\boldsymbol{v}) + w_s E_s(\boldsymbol{v}) + w_a E_a(\boldsymbol{v}) \tag{7}$$

The fitting term has a similar structure as eq. (5), but here minimizes over the gradient instead of exact value, as states in the paper, the gradient is more robust than the direct value of pre-computed lighting model.

$$E_g(\boldsymbol{v}) = \|\nabla B(\boldsymbol{v}) - \nabla \boldsymbol{I}(\boldsymbol{v})\|_2^2 \tag{8}$$

The error term $E_r(\boldsymbol{v})$ regularizes on the geometry of the objective function, and it is also called smoothness, which means that the neighbour voxels should contain similar distance values as the center voxel.

$$E_r(\boldsymbol{v}) = (\Delta \tilde{\boldsymbol{D}}(\boldsymbol{v}))^2 \tag{9}$$

The error term $E_s(\boldsymbol{v})$ stabilizes the refined geometry to avoid influences from noises by constraining the output reconstruction to be close to the unrefined input.

$$E_s(\boldsymbol{v}) = (\tilde{\boldsymbol{D}}(\boldsymbol{v}) - \boldsymbol{D}(\boldsymbol{v}))^2 \tag{10}$$

The fourth term regularizes on the albedo. To do so, it requires the voxels which have similar chromaticity $\Gamma(\boldsymbol{v})$) should hold similar albedo values $\boldsymbol{a}(\boldsymbol{v})$.

$$E_a(\boldsymbol{v}) = \sum_{\boldsymbol{u} \in N_{\boldsymbol{v}}} \phi(\Gamma(\boldsymbol{v}) - \Gamma(\boldsymbol{u})) \cdot (\boldsymbol{a}(\boldsymbol{v}) - \boldsymbol{a}(\boldsymbol{u}))^2 \tag{11}$$

4. Then steps 2 and 3 are repeated several times according to the resolution difference between the depth frame and the color frame. The coarsest level map to the precision of depth frames, and the final integration maps to the resolution of the color frame.

3.2 Computational efficiency approaches

Another main contribution of this method is the computing efficiency. To achieve fast reconstruction, a tailored Gauss-Newton solver is used as well as some help from several previous methods. Voxel hashing is applied here to support the hierarchical refinement; block-based PCG solver is used to speed up the linearized optimization convergence on the GPU.

4 Experiments and results

The paper tested its volumetric shading-based refinement on several datasets including different object scales and sensor types. The final result can achieve sub-millimeter detail. For example, the statues are constructed with noticeable and reasonable curves and edges compared with the original smoothed depth fusion. As introduced in section 2, it also analyzed some results with previous works [3] and [4]. Compared to mesh-based refinement MESHRef [3], volumetric refinement has a clear advantage in computing time. For a synthetically-generated data, MESHRef needs around 20 minutes to perform refinement while volumetric refinement takes less than 7.8 seconds. For a multi-view stereo dataset, MESHRef requires about 1 hour in contrast to 6.5 seconds from the novel method, and the later one also produces higher quality results. When comparing to work [4] which does fusion after refinement, the reconstruction of refinement after fusion contains much more sharpness on the hair and eye region of the Augustus statue, which they used for the comparison.

5 Discussion of results

The results and comparisons are persuasive. However, there is an underlying assumption which states that the "ground truth" obtained from a laser scanner is fully reliable. Only when the assumption holds, all the numerical and qualitative comparisons with ground truth are plausible. Otherwise, the experiments show nothing other than which method can produce more visual appealing results. A better solution is that the author could give more information about the precision of laser scanner. If the error of reconstruction stays near the range of the precision, the comparisons of two similar results may not be able to generate any reasonable conclusion. It may exist ways to acquire true ground truth for a real scene, and yet this is still inevitable problem in 3D reconstruction.

6 Summary

In general, the volumetric shading-based refinement can achieve sub-millimeter fine-scale reconstruction with commodity sensors in only a few seconds. Up to the publish date, it was the first method which could produce such impressive results with regular setups. Nevertheless, it has some limitations. Most important thing is that the efficient computation may not hold for large-scale scene as it takes the whole sequence as the input instead of periodically update the reconstruction. And the method assumes the reconstructed surfaces are Lambertian surface. This assumption lightens the burden of lighting estimation, but also restricts the ability of reconstructing objects with non-Lambertian surfaces, such as specular reflection surfaces. In real world, however, non-Lambertian surface takes a great amount of scenes, typically in man-made world.

References

[1] Robert Maier, Kihwan Kim, Daniel Cremers, Jan Kautz, and Matthias Nießner. Intrinsic3d: High-quality 3d reconstruction by joint appearance and geometry optimization with spatially-varying lighting. *CoRR*, abs/1708.01670, 2017.

[2] Richard A. Newcombe, Shahram Izadi, Otmar Hilliges, David Molyneaux, David Kim, Andrew J. Davison, Pushmeet Kohli, Jamie Shotton, Steve Hodges, and Andrew Fitzgibbon. KinectFusion: Real-time dense surface mapping and tracking. In *Proceedings of the 2011 10th IEEE International Symposium on Mixed and Augmented Reality*, ISMAR '11, pages 127–136, Washington, DC, USA, 2011.

[3] Chenglei Wu, Kiran Varanasi, Yebin Liu, Hans-Peter Seidel, and Christian Theobalt. Shading-based dynamic shape refinement from multi-view video under general illumination. pages 1108–1115, 11 2011.

[4] Chenglei Wu, Michael Zollhöfer, Matthias Nießner, Marc Stamminger, Shahram Izadi, and Christian Theobalt. Real-time shading-based refinement for consumer depth cameras. *ACM Transactions on Graphics*, 33:1–10, 11 2014.

[5] Michael Zollhöfer, Angela Dai, Matthias Innman, Chenglei Wu, Marc Stamminger, Christian Theobalt, and Matthias Nießner. Shading-based refinement on volumetric signed distance functions. ACM, 2015.