Report on BundleFusion:Real-time Globally Consistent 3D Reconstruction using On-the-fly Surface Re-integration, Dai et al, 2017

Iulia-Otilia Mustea

Technische Universität München

Abstract

High quality 3D scanning has a multitude of challenges, including drift in pose estimation and the lack of real-time usage. Recent online methods demonstrated good results, but BundleFusion by Dai et al. outperforms state of the art systems by speed and quality of results. BundleFusion is a real-time end-to-end reconstruction framework, which is based on a local-to-global pose estimation strategy, by considering the current and previous RGB-D input frames. Also, a parallel optimization technique is being used, based on sparse-to-dense features and geometric and photometric matching.

The authors' approach estimates bundle adjusted poses in real-time, supports tracking by ensuring global consistency, re-localization, while dealing with loop closures and re-estimating the 3D model in real-time.

1 Introduction

Applications in robotics, augmented reality and gaming opened the need for real-time 3D scanning and reconstruction systems, with integration into high-quality models, that also provide feedback for the user. In order to achieve such results, the requirements needed are:

- For the model to have a continuous surface, there has to be a high quality representation of the surfaces, rather than point clouds (used in [11]).
- There is a need to acquire models of large spaces, while preserving both the global structure and the local accuracy.
- Robust tracking is required in order to deal with pose drift generated problems, such as loop closures, re-localization, while ensuring global consistency.
- There is a need to continuously integrate acquired data and update the 3D model in a real-time manner, according to new pose estimates.

The aim of BundleFusion is to address all the requirements described previously in an end-to-end 3D reconstruction system. At its core, there lies a local-to-global pose optimization technique. By globally correlating the frames, drift related problems such as loop closures are handled continuously, meaning there is no need for an explicit loop closure detection, which reduces significantly the computation time and makes it robust. Also, a key component of the BundleFusion approach is the parallelized sparse-to-dense optimization, where sparse features, photometric and geometric transformations are used for the coarse-to-fine scale alignment. This technique facilitates both a consistent global structure and high-accurate local surface details. The RGB-D reintegration strategy of on-the-fly-scene updates ensures that the method can be used in real time.

2 Related work

Concerning 3D scanning and reconstruction, there has been a lot of work over the past years. While each of the used methods had both strengths and weaknesses, volumetric methods based on truncated signed functions have been lately used for high-quality reconstructions. One recent example is KinectFusion [8], which was the first paper to show that it was possible to scan and

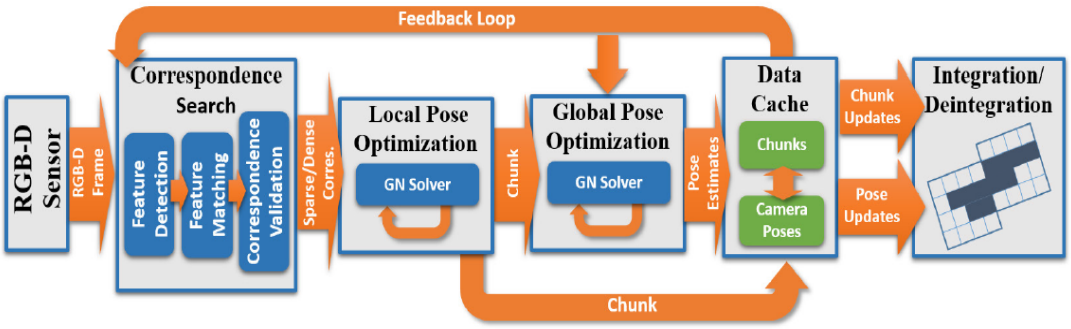
reconstruct 3D objects in real time using a volumetric based approach. Concerning surface representations, ElasticFusion [11] supports a point-based representation, which was proven to limit the scan completeness.

Likewise, most methods require access to all input RGB-D data frames and do offline processing, which don't enable using the method in a real-time manner [2].

Most real time approaches use a frame-to-model iterative closest point (ICP) algorithm [1], which is very efficient,but alignment occurs on a frame to frame basis, which means it can be very easy to accumulate drift over time. To avoid these issues, researchers have tried to use: loop closure detection [10], incremental bundle adjustment [4] or recovery from tracking failures by image or keypoint-based relocalization [5].

3 Method overview

Figure 1: Method overview [3]: BundleFusion takes as input RGB-D streams obtained by using commodity sensors, detects correspondences between input frames and performs a local-to-global pose alignment using sparse to dense correspondences to compute the estimate.



RBG-D streams captured by commodity sensors are used as inputs. The captured frames are then matched against all previous frames, while looking for correspondences between them. Sparse correspondences are found through pairwise Scale-Invariant Feature Transform (SIFT) [7] detection.

For global alignment, BundleFusion performs a sparse-to-dense global pose optimization: firstly, it uses sparse features for a global pose alignment, and then the results are refined through dense photometric and geometric transformations. The used approach is a coarse-to-fine alignment. For the global pose alignment, BundleFusion performs a local-to-global optimization: every consecutive n frames compose a chunk, which is locally optimized. Then, all chunks are combined with respect to each other, thus obtaining a global optimization. Pose alignment is formulated as energy minimization, in which a tailored Gauss-Newton method is used.

The dense scene reconstruction is obtained using a sparse volumetric representation, where on-the-fly scene updates are allowed. In order to update for a new pose, the RGB-D image of the old pose can be removed by de-integrating and re-integrating it at a new pose. Thus, BundleFusion scans and reconstructs a consistent 3D model.

The input to BundleFusion is the RGB-D stream S = $f_i = (C_i, D_i)$, where C_i and D_i represent the color and depth of the i-th frame. There are taken into account only rigid camera transforms T_i, where transformations $T_i(p) = R_i p + t_i$ (rotation R_i, translation t_i) map from the camera coordinates to the world space coordinates.

3.1 Feature Correspondence Search

In BundleFusion, the input frames are pairwise searched for feature detection, feature matching and correspondence filtering steps. For each new input frame, Scale-Invariant Feature Transform (SIFT) [7] features are detected and matched to the features of all previous frames. SIFT is used

as it deals with the transforms encountered during scanning: scaling, rotation and translation. The aim is high precision, since any misalignment will affect the later global pose optimization.

Then, potential matches between each pair of frames are filtered to remove any outliers, in order to produce a list of valid pairwise correspondences, which are later needed for as input the global pose alignment. Thus, BundleFusion uses correspondence filtering techniques.

3.1.1 Key Point Correspondence Filter

First, the authors apply the key point correspondence filter, where they greedily look for low distance correspondences that have a consistent rigid transform between them. They calculate the distance as Root Mean Square Deviation (RMSD) using the Kabsch algorithm [6].

3.1.2 Surface Area Filter

Secondly, they check if the area spanned by the features is large enough (i.e. if there is enough overlap between frames), as correspondences over small sizes are prone to ambiguity. If the areas spanned by 2 sets of correspondences are too small, the set of correspondences is considered ambiguous and discarded.

3.1.3 Dense Verification

Finally, they apply a dense verification where the goal is to look for a high re-projection error, when they project the pixels from one frame into another, under the transform induced by the correspondence set.It is a geometric and phometric verification step. The total re-projection error from f_i to f_j is [3]:

$$E_r(f_i, f_j) = \sum_{x,y} \|T_{ij}(p_{i,x,y}) - q_{j,x,y}\|_2, \tag{1}$$

where $p_{i,x,y} = p_i^{low}(x, y)$ and $q_{j,x,y} = p_j^{low}(\pi^{-1}(T_{ij}p_{i,x,y}))$

Thus, matches between frames f_i and f_j are invalidated in the case of excessive re-projection error. If all the checks are passed after applying the correspondence filtering, then they are added to a valid set, which is later used for pose optimization.

3.2 Hierarchical Optimization

BundleFusion applies a hierarchical optimization strategy. The input sequence of RGB-D data stream is split into chunks of 11 consecutive frames. On the lowest level, alignments within a chunk are optimized. On the higher level, chunks are globally aligned against each other, using representative keyframes per chunk.

The goal of local pose optimization is to get the best intra-chunk alignments. Therefore, valid correspondences are searched pairwise between all frames of the chunk, and then a sparse-to-dense energy minimization approach is applied.

A per-chunk keyframe is defined as the RGB-D data of the chunk's first frame. The keyframes are then used for the global inter-chunk pose optimization. Correspondence search and filtering between global keyframes is similar to that within a chunk, but on the level of all keyframes and their feature sets. The global pose optimization computes the best global alignments for the set of all keyframes of the chunks, therefore it is aligning all chunks globally. Again, the sparse-to-dense energy minimization approach will be applied. The energy optimization technique will be detailed next.

3.3 Pose Alignment as Energy Optimization

For the pose optimization, a sparse-to-dense approach is used, where the energy is a linear combination between the sparse energy term and dense energy term [3]:

$$E_{align}(X) = w_{sparse}E_{sparse} + w_{dense}E_{dense}(X). \tag{2}$$

The sparse correspondences give us a good sense of the global structure, but it cannot be that precise, and the dense energy term is more accurate, but it has a small basin of convergence. Thus, w_{dense} is linearly increased; this allows the sparse term to first find a good global structure, which is then refined with the dense term, therefore achieving coarse-to-fine alignment.

The sparse term, a traditional bundle adjustment, helps with minimizing the sum of distances between the space positions over all correspondences between all pairs of frames. The idea is to seek the best rigid transformations such that the Euclidean distance over all the feature matches is minimized.

$$E_{sparse}(X) = \sum_{i=1}^{|S|} \sum_{j=1}^{|S|} \sum_{(k,l)\epsilon C(i,j)} \|T_i p_{i,k} - T_j p_{j,l}\|_2^2. \tag{3}$$

The dense energy term is a linear combination of depth and color alignment. The dense photometric and geometric constraints are used for fine scale alignment.

$$E_{dense}(X) = w_{photo} E_{photo} + w_{geo} E_{geo}(X). \tag{4}$$

For the photometric term [3], the error of the gradient I_i of the luminance of C_i (i.e. color of the i-th frame) is evaluated to gain robustness against lightning changes.

$$E_{photo}(X) = \sum_{(i,j)\epsilon E} \sum_{k=0}^{|I_i|} \|I_i(\pi(d_{i,k})) - I_j(\pi(T_j^{-1} T_i d_{i,k}))\|_2^2. \tag{5}$$

For the geometric term [3], pixels are taken from one frame, projected into another frame, and then a point to plane distance is computed .

$$E_{geo}(X) = \sum_{(i,j)\epsilon E} \sum_{k=0}^{|D_i|} [n_{i,k}^T (d_{i,k} - T_i^{-1} T_j \pi^{-1}(D_j(\pi(T_j^{-1} T_i d_{i,k}))))]^2. \tag{6}$$

Thus, a coarse-to-fine alignment method is applied.

3.4 Fast and Robust Optimization Strategy

The main part of BundleFusion, the global pose alignment, is actually a non-linear least squares problem in the extrinsic camera parameters. Since the goal is to use the algorithm in real-time, an optimization strategy was needed. The approach is based on the Gauss-Newton method, which requires only the first derivative. The goal is to find the best parameter X^* by minimizing the non-linear least squares problem as in the following equation:

$$X^* = \arg\min_X E_{align}(X). \tag{7}$$

Gauss-Newton is applied with a linear approximation using a first-order Taylor expansion, thus obtaining a system of linear equations. To solve the system, the authors used a Preconditioned Conjugate Gradient (PCG) solver with Jacobi preconditioner, applying a GPU-based parallel approach. By using this approach, the optimization time for the global pose alignment was significantly reduced.

3.5 Dynamic 3D Reconstruction

The final part of the reconstruction is updating the volumetric scene representation based on the optimized camera poses, using integration and de-integration techniques. The scene geometry is constructed by combing the RGB-D data into a truncated signed distance representation (TSDF), which is defined over a volumetric set of voxels. The authors use the sparse volumetric voxel hashing approach by Nießner et al. [9] to scale the reconstruction, since empty space can be discarded.

The frames can be both integrated (i.e.added) and de-integrated (i.e.removed) from the TSDF, which occur as following:

- Integration of a depth frame D_i occurs as following, where each voxel is updated [3] by:

$$D'(v) = \frac{D(v)W(v) + w_i(v)d_i(v)}{W(v) + w_i(v)}, W'(v) = W(v) + w_i(v) \tag{8}$$

- De-integration of a frame is the reversed operation, thus each voxel is updated [3] by:

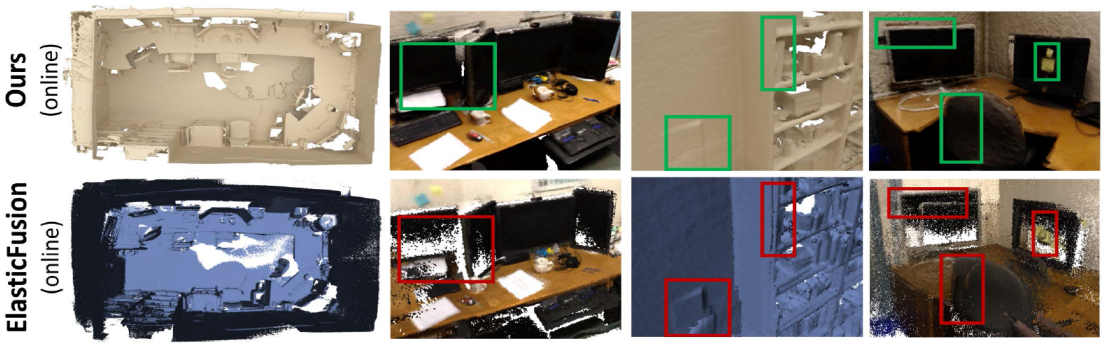
$$D'(v) = \frac{D(v)W(v) - w_i(v)d_i(v)}{W(v) - w_i(v)}, W'(v) = W(v) - w_i(v) \tag{9}$$

Therefore, update a new pose in the reconstruction is done by de-integrating the frame from the original pose and integrating it into a new pose.

4 Results

For real-time scanning, authors used a commodity sensor on an iPad, then the captured RGB-D stream was send to a desktop machine that runs the algorithm. The completess of the large-scale indoor scenes, the alignment without camera drift, the high quality of the geometry is also shown in Fig.2. In comparison with KinectFusion [8], BundleFusion handles loop closures, can recover from tracking failures and reduces drift, which most ICP frame-to-model methods cannot minimize.

Figure 2: Comparison [3]: BundleFusion outperforms ElasticFusion [11] in terms of scan quality, accuracy and completeness.



In terms of performance, authors used for computation two GPUs - NVIDIA GEForce GTX Titan X and GTX Titan Black, which helps the global dense optimization run in less than 500 ms.

In terms of limitations, misalignments can occur due to the SIFT matches, which can be off by a few pixels.Treating the keypoints globally would require much more computational time. Also, live scanning is available just up to 14 minutes, due to the hardware configuration. For longer sequences, more hierarchy levels would be needed.

5 Conclusion

BundleFusion is a real-time 3D scanning and reconstruction approach with commodity RGB-D sensors. It optimizes for all the poses globally in real-time using pairwise SIFT features of RGB-D input frames, enabling a parallel non-linear pose optimization over both sparse and dense features. The optimized poses update the scene on-the-fly through integration and de-integration. Thus, BundleFusion's output is a large-scale consistent and accurate 3D model, whose real-time reconstruction impresses by quality .

References

[1] P.J. Besl and N.D. McKay. A method for registration of 3-d shapes. In *IEEE Trans.PAMI 14,2*, pages 127–136, 1992.

[2] Sungjoon Choi, Qian-Yi Zhou, and Vladen Kotltun. Robust Reconstruction of Indoor Scenes. In *Proc. CVPR*, June 2015.

[3] Angela Dai, Matthis Nießner, Michael Zollhöfer, Shahram Izadi, and Christian Theobalt. BundleFusion: Real-Time Globally Consistent 3D Reconstruction Using On-the-Fly Surface Reintegration. In *ACM Transactions on Graphics (TOG) 36,3*, 2017.

[4] Nicola Fioraio, Jonathan Taylor, Andrew Fitzgibbon, Luigi Di Stefano, and Shahram Izadi. Large-scale and Drift-Free Surface Reconstruction Using Online Subvolume Registration. In *Proc.CVPR*, June 2015.

[5] Ben Glocker, Jamie Shotton, Antonio Criminisi, and Sharam Izadi. Real-time RGB-D camera Relocalization via Randomized Ferns for Keyframe Encoding. In *TVCG 21,5*, pages 571–583, 2015.

[6] Wolfgang Kabsch. A solution for the best rotation to relate two sets of vectors. In *Acta Crystallographica Section A: Crystal Physics, Diffraction, Theoretical and General Crystallography 32,5*, pages 992–923, 1976.

[7] David G. Lowe. Distinctive Image features from Scale-Invariant Keypoints. In *ICJV 60*, pages 91–100, 2004.

[8] 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.

[9] Matthis Nießner, Michael Zollhöfer, Shahram Izadi, and M. Stamminger. Real-time 3D Reconstruction at Scale using Voxel Hashing. In *ACM TOG*, 2013.

[10] Thomas Whelan, Michael Kaess, John J Leonard, and John McDonald. Deformation-based loop closure for large scale dense rgb-d slam. In *Proc.IROS.IEEE*, pages 548–555, 2013.

[11] Thomas Whelan, Stefan Leutenegger, Renato F. Salas-Moreno, Ben Glocker, and Andrew J. Davison. ElasticFusion: Dense SLAM Without A Pose Graph. In *Robotics: Science and Systems 2015*, 2015.