Large-Scale and Drift-Free Surface Reconstruction Using Online Subvolume Registration

Nicola Fioraio[1], Jonathan Taylor[2], Andrew Fitzgibbon[2], Luigi Di Stefano[1], Shahram Izadi[2]
[1]Department of Computer Science and Engineering, University of Bologna. [2]Microsoft Research Cambridge.

Much recent progress has been made in the development of real-time, dense surface reconstruction algorithms that work with a single depth camera, such as the Microsoft Kinect. KinectFusion [3] demonstrated high-quality scanning of small environments, subsequently extended to large-scale reconstructions [2, 4]. A major barrier to complete reconstruction is error accumulation during sequential camera pose estimation. When exploring large environments, this sensor "drift" often corrupts the final reconstruction with artifacts and clear misalignments (see center of Fig. 1). Current solutions usually require either RGB data [5, 6], explicit loop closure [5] or an expensive off-line global optimization step [6]. We propose a novel approach which performs real-time camera tracking while performing online model correction, uses depth data only, avoids explicit loop closure detection and executes a full global surface alignment to facilitate fast dense 3D reconstruction.

A key component of our method is the use of the Truncated Signed Distance Function (TSDF) representation. This is expressed as a pair of functions (F, W) such that, for every $\mathbf{u} \in \mathbb{R}^3$, $F(\mathbf{u})$ is the distance from $\mathbf{u}$ to the zero level set of F, while the weighting function $W(\mathbf{u})$ encodes a measure of confidence in the value of F at $\mathbf{u}$. Given a depth image D_t at time t, first we estimate the corresponding camera pose T_t as proposed in [1], then depth measurements are integrated in the GPU memory as

$$F^{\text{new}}(\mathbf{u}) = \frac{F(\mathbf{u})\,W(\mathbf{u}) + \min\left(1, {}^{\Delta_z(\mathbf{u},t)}/_{\delta}\right)}{W(\mathbf{u})+1}\ , \tag{1}$$

$$W^{\text{new}}(\mathbf{u}) = W(\mathbf{u}) + 1\ , \tag{2}$$

where Δ_z is the difference between the voxel position in camera space and the measured depth, while δ is the truncation distance. However, dense estimation of the TSDF requires a large amount of memory which is practical only for small workspaces. Though moving volume approaches [2, 4] allow for virtually unbounded reconstruction, large exploratory sequences introduce drift error in the estimated camera trajectory, leading to gross misalignments and artifacts. In our approach, camera tracking is always performed against a low-drift, high quality, local TSDF produced by the fusion of the last K tracked frames. Specifically, the current tracked frame D_t is integrated into the TSDF and then pushed into a FIFO queue, while the K^{th} oldest frame D_{t-K} gets popped and *eroded* by applying

$$F^{\text{new}}(\mathbf{u}) = \frac{F(\mathbf{u})\,W(\mathbf{u}) - \min\left(1, {}^{\Delta_z(\mathbf{u},t-K)}/_{\delta}\right)}{W(\mathbf{u})-1}\ , \tag{3}$$

$$W^{\text{new}}(\mathbf{u}) = W(\mathbf{u}) - 1\ . \tag{4}$$

At the same time, every K frames the current TSDF is frozen and copied into main memory as a *subvolume*. Camera tracking continues on the GPU, while drift is reduced on the host side through global bundle adjustment and surface alignment of all subvolumes hitherto available.

Global surface alignment is addressed by finding the best rigid-body pose for each subvolume, encoded as a transform V_j for the j'th subvolume (see Fig. 1). Purposely, a cost function is built by establishing a set of correspondences as follows. For each point $\mathbf{p}_i^{(j)}$ on the zero level set of subvolume (F_j, W_j), we define a match $\mathbf{q}_k^{ji}$ with overlapping subvolume (F_k, W_k) by following the direction of the normalized gradient $\widehat{\nabla} F_k$ to obtain

$$\mathbf{q}_k^{ji} = \mathrm{V}_k^{-1}\mathrm{V}_j\mathbf{p}_i^{(j)} - F_k\left(\mathrm{V}_k^{-1}\mathrm{V}_j\mathbf{p}_i^{(j)}\right)\widehat{\nabla}F_k\left(\mathrm{V}_k^{-1}\mathrm{V}_j\mathbf{p}_i^{(j)}\right)\ . \tag{5}$$

The cost function is then the sum of point-to-plane distances

$$\sum_{jik}\left\|\left(\mathbf{p}_i^{(j)} - \mathrm{V}_j^{-1}\mathrm{V}_k\mathbf{q}_k^{ji}\right)\cdot\widehat{\nabla}F_j\left(\mathbf{p}_i^{(j)}\right)\right\|^2\ , \tag{6}$$

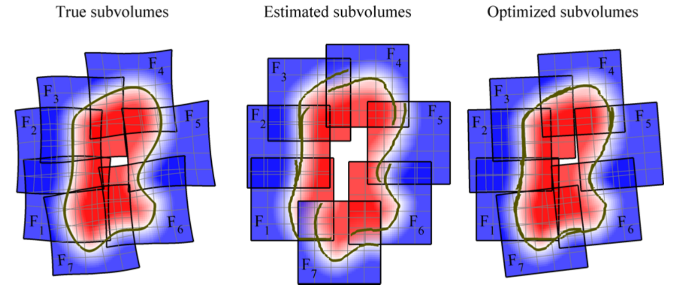

Figure 1: Locally-accumulated TSDFs (left) are shipped from GPU to host, effectively representing the *true* scene by a set of overlapping individually reconstructed subvolumes (center). Noise and drift are reduced by estimating an optimized 6-DOF pose for each subvolume (right).

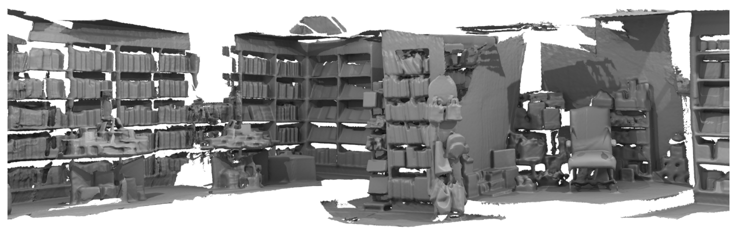

Figure 2: Reconstruction of a bookshop.

and is minimized through non-linear least squares. Given the refined subvolumes' poses, another set of matches is found and a new cost function is built and optimized. This process is repeated until convergence.

At the end of this procedure, we have estimated a 6-DOF pose for each subvolume, but non-rigid deformations still show up as artifacts when extracting surfaces. Instead of computing a global volume by re-sampling subvolumes, we deploy a faster volume blending approach. Accordingly, for each point $\mathbf{u}$ in subvolume (F_j, W_j) we consider the set of overlapping subvolumes $\{(F_k, W_k)\}$ and update the distance function as

$$F_j^{\text{new}}(\mathbf{u}) = \frac{F_j(\mathbf{u})\,W_j(\mathbf{u}) + \sum_k F_k\left(\mathrm{V}_k^{-1}\mathrm{V}_j\mathbf{u}\right)W_k\left(\mathrm{V}_k^{-1}\mathrm{V}_j\mathbf{u}\right)}{W_j(\mathbf{u}) + \sum_k W_k\left(\mathrm{V}_k^{-1}\mathrm{V}_j\mathbf{u}\right)}\ . \tag{7}$$

An exemplar reconstruction of a large environment achieved by our method is shown in Fig. 2). Finally, it is worth pointing out that volume blending is not required by either camera tracking or subvolume registration. Likewise, the result of subvolume optimization is not needed by the camera tracking process, which can thus keep operating in real-time.

[1] Bylow, Sturm, Kerl, Kahl, and Cremers. Real-time camera tracking and 3d reconstruction using signed distance functions. In *RSS*, June 2013.

[2] Chen, Bautembach, and Izadi. Scalable real-time volumetric surface reconstruction. *ACM Trans. on Graphics (TOG)*, 32(4), July 2013.

[3] Newcombe, Izadi, Hilliges, Molyneaux, Kim, Davison, Kohli, Shotton, Hodges, and Fitzgibbon. KinectFusion: Real-time dense surface mapping and tracking. In *ISMAR*, pages 127–136, 2011.

[4] Whelan, McDonald, Kaess, Fallon, Johannsson, and Leonard. Kintinuous: Spatially extended KinectFusion. In *RSS Workshop on RGB-D: Advanced Reasoning with Depth Cameras*, Sydney, Australia, July 2012.

[5] Whelan, Kaess, Leonard, and McDonald. Deformation-based loop closure for large scale dense RGB-D SLAM. In *IROS, IEEE/RSJ Intl. Conf.*, Tokyo, Japan, November 2013.

[6] Zhou and Koltun. Dense scene reconstruction with points of interest. *ACM Trans. on Graphics (TOG)*, 32(4), July 2013.

This is an extended abstract. The full paper is available at the Computer Vision Foundation webpage.