DynamicFusion: Reconstruction and Tracking of Non-rigid Scenes in Real-Time

Richard Newcombe, Dieter Fox, Steve Seitz
Department of Computer Science and Engineering, University of Washington.

3D scanning traditionally involves separate capture and off-line processing phases, requiring very careful planning of the capture to make sure that every surface is covered. In practice, it's very difficult to avoid holes, requiring several iterations of capture, reconstruction, identifying holes, and recapturing missing regions to ensure a complete model. Real-time 3D reconstruction systems like KinectFusion [3] represent a major advance, by providing users with the ability to instantly see reconstructions of the surface geometry; identify regions that remain to be scanned; and provide immediate predictions of the observed surfaces for use in augmented reality. But like all traditional SLAM and dense reconstruction systems, the most basic assumption behind KinectFusion is that the scene is largely *static*. To that end, we introduce DynamicFusion, an approach based on solving for a volumetric flow field that transforms the state of the scene at each time instant into a fixed, canonical frame. In the case of a moving person, for example, this transformation undoes the person's motion, warping each body configuration into the pose of the first frame. Following these warps, the scene is effectively rigid, and standard KinectFusion updates can be used to obtain a high quality, denoised reconstruction. This reconstruction can then be transformed back into the live frame using the inverse map; each point in the canonical frame is transformed to its location in the live frame (see Figures 1,3)

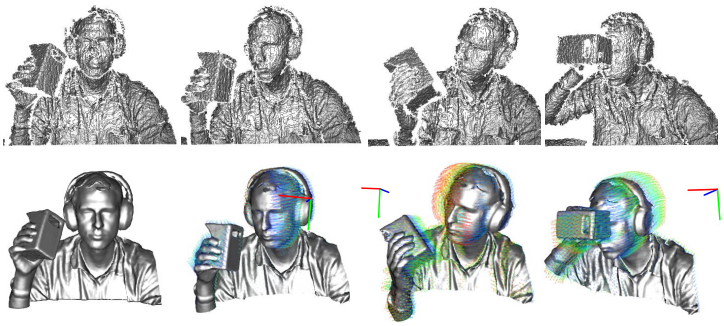

Figure 1: DynamicFusion takes in a stream of noisy depth maps (top) and outputs a dense reconstruction of the moving scene (bottom). Motion trails are shown for correspondences from the last 1 second of motion, with a coordinate frame showing the rigid body component of the scene motion.

Our main insight is that undoing the scene motion to enable fusion of all observations into a single *fixed* frame can be achieved efficiently by computing the inverse map alone. Under this transformation, each canonical point projects along a line of sight in the live camera frame (see Figure 2). Since the optimality arguments of [1] (developed for rigid scenes, and used in the KinectFusion algorithm) depend only on lines of sight, we can generalize their optimality results to the non-rigid case.

While no prior work achieves real-time, template-free, non-rigid reconstruction, there are two categories of closely related work: 1) real-time non-rigid tracking given a prior template e.g., [4], and 2) offline dynamic reconstruction techniques often with multiple cameras e.g., [2].

Algorithm Outline: DynamicFusion decomposes a non-rigidly deforming scene into a latent geometric surface, reconstructed into a rigid canonical space $\mathbf{S} \subseteq \mathbb{R}^3$; and a per frame volumetric warp field that transforms that surface into the live frame. There are three core algorithmic components to the system that are performed in sequence on arrival of each new depth frame:

1. Estimation of the volumetric model-to-frame warp field parameters.
2. Fusion of the live frame depth map into the canonical space via the estimated warp field.
3. Adaptation of the warp-field structure to capture newly added geometry.

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

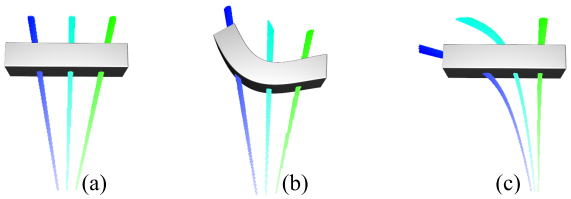

Figure 2: An illustration of how space is deformed by the warp field in the (rigid) canonical frame to transform points into the live (non-rigid) frame. We highlight all of the voxels which project onto three pixels in a live frame of a non-rigidly deforming scene (a). In the corresponding canonical frame the warp field is initialized to the identity transform and all these voxels also lie along straight lines. As the scene deforms in the live frame (b), we can observe how the warp function transforms each point in the canonical frame back to the live frame resulting in bending of the correcponding rays (c).

Specifically, we enable non-rigid tracking and scene reconstruction by extending KinectFusion in two ways. First, we replace the single rigid body camera to world transform that describes the relative scene motion with a dense volumetric warp-field, providing a per point $6D$ transformation that rotates and translates the canonical space into the live frame: $\mathcal{W} : \mathbf{S} \mapsto \mathbf{SE}(3)$. For each canonical point $v_c \in \mathbf{S}$, the transform $\mathcal{W}(v_c)$ warps v_c from canonical space into the live, non-rigidly deformed frame of reference. Second, we generalise the truncated signed distance function (TSDF) fusion approach originally introduced by [1] to operate over non-rigidly deforming scenes (see Figure 2). Given an accurately estimated warp field then the live frame location of each canonical point is known, and the projective signed distance to that live frame point computed (as in KinectFusion). Since all TSDF updates are computed using distances in the camera frame, the non-rigid projective TSDF fusion approach maintains the optimality guarantees for surface reconstruction from noisy observations originally proved for the static reconstruction case in [1].

Figure 3: The first view and final canonical model reconstruction from DynamicFusion results shown in our accompanying video available at http://grail.cs.washington.edu/projects/dynamicfusion.

We believe that DynamicFusion will open up a number of interesting applications of real-time 3D scanning and SLAM systems in dynamic environments.

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

[2] Mingsong Dou, H. Fuchs, and J.-M. Frahm. Scanning and tracking dynamic objects with commodity depth cameras. In *Mixed and Augmented Reality (ISMAR), 2013 IEEE International Symposium on*, 2013.

[3] 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 *Proceedings of the International Symposium on Mixed and Augmented Reality (ISMAR)*, 2011.

[4] Michael Zollhöfer, Matthias Niessner, Shahram Izadi, Christoph Rehmann, Christopher Zach, Matthew Fisher, Chenglei Wu, Andrew Fitzgibbon, Charles Loop, Christian Theobalt, and Marc Stamminger. Real-time Non-rigid Reconstruction Using an RGB-D Camera. *ACM Trans. Graph.*, 33(4), July 2014.