3D Scanning Deformable Objects with a Single RGBD Sensor

Mingsong Dou[1], Jonathan Taylor[2], Henry Fuchs[1], Andrew Fitzgibbon[2], Shahram Izadi[2]
[1]Department of Computer Science, UNC-Chapel Hill. [2]Microsoft Research.

Many existing 3D scanning systems use rigid alignment algorithms and thus require the object or scene being scanned to remain static. In many scenarios such as reconstructing humans, particularly children, and animals, nonrigid movement is *inevitable*. Recent work on nonrigid scanning are constrained by relying on specific user motion (e.g. [4]); requiring multiple cameras (e.g. [1]); using a static pre-scan as template prior (e.g. [5]); or nonrigidly aligning partial static scans (e.g. [2]).

To address these issues we present a new 3D scanning system for arbitrary scenes, based on a single sensor, which allows for large deformations during acquisition. Further, our system avoids the need for any static capture, either as a template prior or for acquiring initial partial scans. Our goal is to combine a sequence of depth images, each representing a noisy and incomplete scan of the object of interest, into a high quality and complete 3D model. A major problem to address is that of *drift* in which the error in the alignment between subsequent scans accumulates quickly and the scan does not close seamlessly. We address this by automatically detecting these "loop closures".However, dealing with such loop closures only allows the error to be distributed equally over the loop instead of actually minimizing the error. We, therefore, also perform a dense nonrigid bundle adjustment to simultaneously optimize the latent 3D shape and the nonrigid deformations needed to minimize this error in each frame. Our experiments show that this bundle adjustment gives improved data alignment and a high quality final model. Figure 1 illustrates the system pipeline.

The first phase of our algorithm begins by preprocessing a RGBD sequence into a shorter sequence of N high quality, but only partial, scans $\{\mathcal{V}_i\}_{i=1}^N$ of the object of interest. Each partial scan $\mathcal{V}_i$ is a triangular mesh obtained by fusing a small contiguous set of frames using the method of [1]. Ultimately, we want each partial scan $\mathcal{V}_i$ to be explained by a plausible deformation of a single latent mesh $\mathcal{V} = \{v_m\}_{m=1}^M$ representing the latent 3D structure of the object being scanned. We parameterize this non-rigid deformation using the embedded deformation (ED) model [3]. In this model, a set of K "ED nodes" are sampled throughout the mesh at a set of fixed locations $\{\mathbf{g}_k\}_{k=1}^K \subseteq \mathbb{R}^3$ and each vertex m is "skinned" to the ED nodes by a set of fixed weights $\{w_{mk}\}_{k=1}^K \subseteq [0,1]$, which are set as a function of the geodesic distance between vertex m and ED node k. To each ED node is associated a 3D affine transformation $\{A_k, \mathbf{t}_k\}$. Under this model, the deformed location of vertex $\mathbf{v}_m$ using the parameter set $G = \{R,T\} \cup \{A_k, \mathbf{t}_k\}_{k=1}^K$ is

$$ED(\mathbf{v}_m;G) = R\sum_{k=1}^{K} w_{mk}\left[A_k(\mathbf{v}_m-\mathbf{g}_k)+\mathbf{g}_k+\mathbf{t}_k\right]+T\ , \tag{1}$$

where R,T represent global rigid transformation.

We fit this model using a bundle adjustment (BA) technique to refine the latent mesh $\mathcal{V}$ as to explain all the data summarized in the partial scans $\{\mathcal{V}_i\}_{i=1}^N$. For each data point $\mathbf{v}_m^i$ in segment $\mathcal{V}_i$, we expect ED parameters G_i to deform it so that is consistent with the latent mesh. We thus employ an energy function designed to encourage a small distance between $ED(\mathbf{v}_m^i;G_i)$ and the latent surface, and for the normal to match. This term is

$$E_{\text{data}}(\mathcal{V}) = \sum_{i=1}^{N}\min_{G_i}\sum_{m=1}^{M_i}\min_{\mathbf{u}}\ E_{\text{point}}(\mathbf{v}_m^i;G_i,\mathbf{u},\mathcal{V}) + E_{\text{normal}}(\mathbf{n}_m^i;G_i,\mathbf{u},\mathcal{V})$$

where

$$\begin{aligned} E_{\text{point}}(\mathbf{v};G,\mathbf{u},\mathcal{V}) &= \lambda_{\text{data}}\|ED(\mathbf{v};G)-S(\mathbf{u};\mathcal{V})\|^2 \\ E_{\text{normal}}(\mathbf{n};G,\mathbf{u},\mathcal{V}) &= \lambda_{\text{normal}}\|ED(\mathbf{n};G)-S^{\perp}(\mathbf{u};\mathcal{V})\|^2\ . \end{aligned}$$

$S(\mathbf{u};\mathcal{V})$ and $S^{\perp}(\mathbf{u};\mathcal{V})$ represent the point and normal respectively of a coordinate $\mathbf{u}$ (*i. e.* a triangle index and barycentric coordinate) on the latent mesh surface. In addition to the above data term, we add terms to regularize both the nonrigid motion and latent mesh.

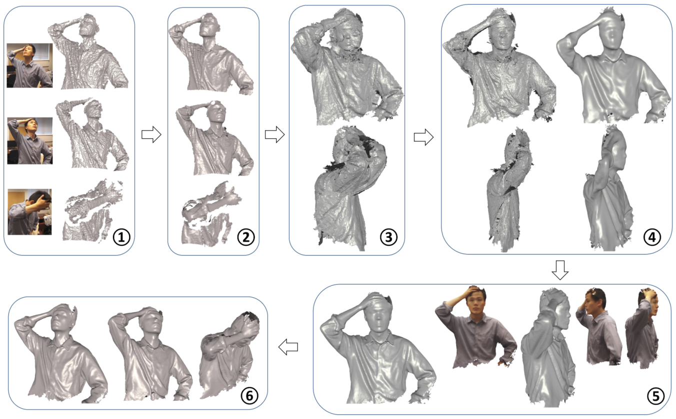

Figure 1: Scanning pipeline: ① input depth and color; ② partial scans; ③ coarse-aligned scans; ④ Loop closure (LC): aligned scans (left) and fused latent mesh (right); ⑤ mesh after bundle adjustment; ⑥ meshes deformed to every frame. The input sequence has around 400 frames which are fused into 40 partial scans. These partial scans are then consecutively placed in the reference pose to achieve the coarse alignment. Next, loop closures are detected and alignment is refined; all the LC-aligned scans are fused volumetrically to get the LC-fused surface which serves as the initial for the following bundle adjustment stage. The final model can be deformed back to each frame to reconstruct the whole sequence.

The final cost function can be optimized with a standard nonlinear least squares solver, however this first requires a reasonable initialization of the parameters. To this end, we first obtain a coarse alignment for all partial scans $\{\mathcal{V}_i\}_{i=1}^N$ by nonrigidly aligning adjacent frame pairs and consecutively deforming them to the reference pose (*i. e.*, pose of $\mathcal{V}_1$). Naturally, the error in the alignment step accumulates, making the deformation parameter sets more and more unreliable as i increases. We assume, however, that our sequence includes a loop closure and thus there should be some later segments that could match reasonably well with earlier segments. We identify such pairs and establish rough correspondences between them. With these loop closing correspondences extracted, we use Li *et al.*'s algorithm [2] to re-estimate the ED graph parameters $\mathcal{G} = \{G_i\}_{i=1}^N$. We find that that these parameters are of sufficient quality that our bundle adjustment procedure converges to a high quality reconstruction.

In contrast to previous systems, a wider range of deformations can be handled, and more geometry details are recovered through the bundle adjustment stage. Some limitations remain, however. First, although complex scene topologies can be handled, the topology is restricted to be constant throughout the sequence. The computational cost is also high. Our CPU-implmented bundle adjustment algorithm takes several hours to converge.

[1] M. Dou, H. Fuchs, and J.-M. Frahm. Scanning and tracking dynamic objects with commodity depth cameras. In *Proc. ISMAR*, 2013.

[2] H. Li, E. Vouga, A. Gudym, L. Luo, J. T. Barron, and G. Gusev. 3d self-portraits. *ACM Trans. Graph.*, 32(6):187, 2013.

[3] R. W. Sumner, J. Schmid, and M. Pauly. Embedded deformation for shape manipulation. In *SIGGRAPH*, 2007.

[4] J. Tong, J. Zhou, L. Liu, Z. Pan, and H. Yan. Scanning 3D full human bodies using Kinects. *TVCG*, 18(4):643–650, 2012.

[5] M. Zollhöfer, M. Nießner, S. Izadi, C. Rehmann, C. Zach, M. Fisher, C. Wu, A. Fitzgibbon, C. Loop, C. Theobalt, and M. Stamminger. Real-time non-rigid reconstruction using an rgb-d camera. *ACM Transactions on Graphics (TOG)*, 33(4), 2014.

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