Coloured signed distance fields for full 3D object reconstruction

Wadim Kehl[1]
kehl@in.tum.de

Nassir Navab[1]
navab@in.tum.de

Slobodan Ilic[2]
slobodan.ilic@siemens.com

[1] CAMP Chair
Computer Science Department
TU Munich, Germany

[2] Siemens AG
Research & Technology Center
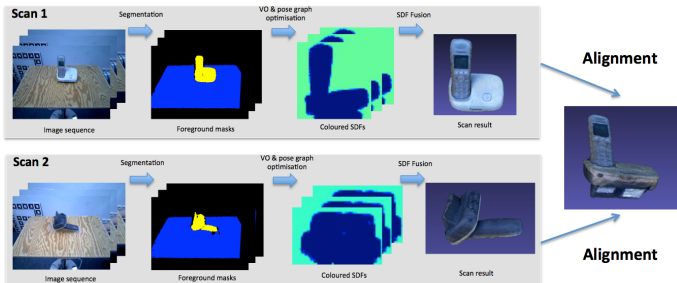Munich, Germany

Figure 1: We take foreground-masked sequences, pose-optimise and fuse each of them and eventually align them to one coherent 3D model.

We propose a full 3D object reconstruction framework with a RGB-D sensor, requiring no marker boards and allowing for objects to be displaced during scanning. The proposed framework consists of three stages and provides a novel fusion and registration procedure for coloured signed distance fields (CSDFs) resulting in complete 3D models with high fidelity. It is suitable for a large variety of objects and outperforms the state-of-the-art both in terms of visual quality and geometrical accuracy.

The first step in our pipeline is the camera trajectory estimation via RGB-D visual odometry [5] similar to [2]. The goal is to compute the rigid-body movement $\Xi \in SE(3)$ of the camera between two consecutive foreground-masked sensor pairs $[I_0, D_0], [I_1, D_1]$ by minimising

$$E(\Xi) = \int_{\Omega_2} [I_1(w_\Xi(x)) - I_0(x)]^2 dx$$

with a warp function $w_\Xi : \Omega_2 \to \Omega_2$ defined via the depth maps as $w_\Xi(x) = \pi_{D_1}(\Xi \cdot \pi_{D_0}^{-1}(x))$. We move the support surface while collecting keyframes along the way and eventually refine the trajectory globally with a pose-graph optimisation after loop closure detection.

After one full scan and pose refinement, we refer to our final result as a hemisphere $\mathcal{H} = \{(I_i, D_i, P_i)_i\}$ consisting of masked sensor pairs and poses. We create a 3D model ϕ by fusing the data, analogously to [1, 3], into a CSDF in a variational fashion with an approximate L^1 minimisation. We cast our data into volumetric geometry fields $f_i : \Omega_3 \subset \mathbb{R}^3 \to \mathbb{R}$ and colour fields $c_i : \Omega_3 \to [0,1]^3$ and we seek the minimisers of the functional

$$\mathcal{E}(u,v) = \int_{\Omega_3} [\mathcal{D}(\mathbf{f}, \mathbf{w}, \mathbf{c}, u, v) + \alpha \mathcal{S}(\nabla u) + \beta \mathcal{S}(\nabla v)] \ d\mathbf{x}$$

with a data term $\mathcal{D}$ that strives to uphold the solution's fidelity to all the observations $\mathbf{f} = \{f_1, ..., f_n\}, \mathbf{c} = \{c_1, ..., c_n\}$ and two weighted regularisers $\mathcal{S}(\nabla u)$ and $\mathcal{S}(\nabla v)$. In contrast to the original work [4], which only fuses the geometrical fields, we also include colour information into the formulation and solve simultaneously for both.

A suitable data term for many vision problems usually involves an outlier-robust L^1-norm whereas for regularisation purposes the total variation (TV) of the function is often employed:

$$\mathcal{D}(\mathbf{f}, \mathbf{w}, \mathbf{c}, u, v) = \frac{1}{\varepsilon + \sum_i w_i} \sum_i w_i \cdot (|u - f_i| + |v - c_i|) \ , \ \mathcal{S}(\nabla u) = |\nabla u|.$$

Due to the problematic aspect of solving such energies, specific minimisation schemes are employed (e.g. a ROF-variant or (iterated) primal-dual solutions). An alternative has been proposed in [4] where the problematic terms have been replaced with a smooth $epsL^1$ approximation $\Gamma(x) := \sqrt{x^2 + \varepsilon}$. We define it similarly as

$$\mathcal{D}(\mathbf{f}, \mathbf{w}, \mathbf{c}, u, v) = \Gamma(\sum_i w_i)^{-1} \sum_i w_i \cdot (\Gamma(u - f_i) + \Gamma(v - c_i)) \ , \ \mathcal{S}(\nabla u) = \Gamma(|\nabla u|)$$
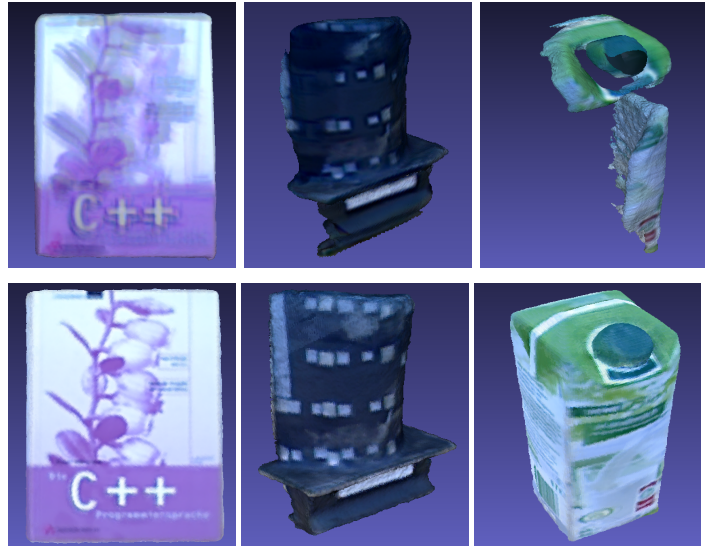

Figure 2: KinectFusion (top) vs. our approach (bottom). We recover richer texture as well as geometry, even for visually poor objects.

where we regard the weighted approximate absolute differences together with an additional normalisation factor and an approximate TV-regulariser.

Usually, one such scan does not expose the full geometry of the object. To this end, we propose to create multiple scans of the same object but placed differently in order to reveal hitherto unseen parts, thus acquiring multiple hemispheres $\mathcal{H}_j$. Then the transformations Ξ_j that map the models from all hemispheres to the first one $\mathcal{H}_0$ need to be determined. In order to retrieve those Ξ_j, we use the reconstructed models ϕ_j and align them automatically using a dense approximate-L^1 registration framework:

$$\mathcal{E}(\Xi)_{L^1} = \int_{\Omega_3} \Gamma(\phi_0(\mathbf{x}) - \phi_j(\Xi(\mathbf{x}))) \ d\mathbf{x}.$$

We compared our method to a commercial state-of-the-art KinectFusion implementation on eight real-life objects. Even though KinectFusion performed well, it failed for some of the objects due to poor geometry leading to tracking failure and supplied only mediocre results in terms of texture. For two models ground-truth data was available and was used to measure the geometrical error of the reconstructions. We show that we tremendously boost the geometrical and textural fidelity for all scanned objects due to the pose graph optimisation and the L^1 sensor fusion.

[1] Brian Curless and Marc Levoy. A volumetric method for building complex models from range images. In *SIGGRAPH*, 1996.

[2] Maria Dimashova, Ilya Lysenkov, Vincent Rabaud, and Victor Eruhimov. Tabletop Object Scanning with an RGB-D Sensor. *ICRA Workshop*, 2013.

[3] Richard A. Newcombe, Andrew J. Davison, Shahram Izadi, Pushmeet Kohli, Otmar Hilliges, Jamie Shotton, David Molyneaux, Steve Hodges, David Kim, and Andrew Fitzgibbon. KinectFusion: Real-time dense surface mapping and tracking. In *ISMAR*, 2011.

[4] Christopher Schroers, Henning Zimmer, Levi Valgaerts, Oliver Demetz, and Joachim Weickert. Anisotropic Range Image Integration. In *Pattern Recognition, LNCS*, 2012.

[5] Frank Steinbrucker, Jurgen Sturm, and Daniel Cremers. Real-time visual odometry from dense RGB-D images. In *ICCV Workshop*, 2011.