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

ISAKOV Vladimir, LASBLEIS Alexandre

University of Amsterdam

26/04/17

Introduction

- Paper from Richard A. Newcombe, Dieter Fox, Steven M. Seitz.
- 2015 best paper of the year award

Introduction

- First SLAM system capable of reconstructing non-rigidly deforming scene in real time. (`https://youtu.be/i1eZekcc_lM?t=29`)
- Mainly inspired by Kinect Fusion: Real-time 3D Reconstruction and Interaction Using a Moving Depth Camera.
- Transform scene to *canonical frame* to undo the object motion.
- Do not require template or any prior scene model and use a single depth camera.

Related works 1/3

Real-time non-rigid template tracking (focuses on human body parts):

- faces - H. Li, J. Yu, Y. Ye, and C. Bregler. Realtime Facial Animation with On-the-fly Correctives.
- hands - I. Oikonomidis, N. Kyriazis, and A. Argyros. Efficient model-based 3D tracking of hand articulations using Kinect.
- complete bodies - J. Taylor, J. Shotton, T. Sharp, and A. Fitzgibbon. The Vitruvian manifold: Inferring dense correspondences for oneshot human pose estimation.
- general articulated objects - T. Schmidt, R. Newcombe, and D. Fox. DART: Dense Articulated Real-Time Tracking. Proceedings of Robotics: Science and Systems.

Related works 2/3

Offline simultaneous tracking and reconstruction of dynamic scenes(1/2):

- extended ICP for small non-rigid deformations - B. Brown and S. Rusinkiewicz. Non-Rigid Range-Scan Alignment Using Thin-Plate Splines.
- pairwise 3D shape and scan alignment over larger deformations - W. Chang and M. Zwicker. Range Scan Registration Using Reduced Deformable Models.
- embedded deformation graphs - R. W. Sumner, J. Schmid, and M. Pauly. Embedded deformation for shape manipulation.

Related works 3/3

Offline simultaneous tracking and reconstruction of dynamic scenes(2/2):

- quasi-rigid reconstruction - M. Zeng, J. Zheng, X. Cheng, and X. Liu. Templateless Quasi-rigid Shape Modeling with Implicit Loop-Closure.
- non-rigid shape denoising - Q. Zhang, B. Fu, M. Ye, and R. Yang. Quality Dynamic Human Body Modeling Using a Single Low-cost Depth Camera.

Method overview 1/2

DynamicFusion decomposes a non-rigidly deforming scene into a latent geometric surface, reconstructed into a rigid canonical space $S \in R^3$; and a per frame volumetric warp field that transforms that surface into the live 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

Method overview 2/2

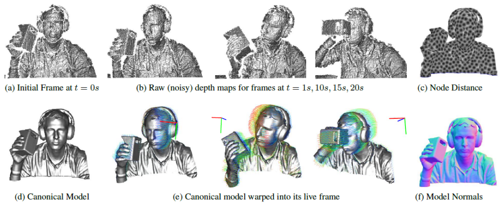

(a) Initial Frame at $t = 0s$
(b) Raw (noisy) depth maps for frames at $t = 1s, 10s, 15s, 20s$
(c) Node Distance
(d) Canonical Model
(e) Canonical model warped into its live frame
(f) Model Normals

Figure 1: Method overview

Dense Non-Rigid Warp Field

Represent dynamic scene motion through a volumetric warp-field, providing a per point 6D transformation. The warp field is constructed as a set of sparse 6D transformation nodes that are smoothly interpolated through a k-nearest node average in the canonical frame.
Dense parametrization of the warp function is infeasible, use dual-quaternion blending DQB to define the warp function:

$$W(x_c) = SE3(DQB(x_c))$$

$$DQB(x_c) = \frac{\sum_{k \in N(x_c)} w_k(x_c) q_{kc}}{\|\sum_{k \in N(x_c)} w_k(x_c) q_{kc}\|}$$

Volumetric warp function:

$$W(x_c) = T_{lw} SE3(DQB(x_c))$$

Dense Non-Rigid Surface Fusion 1/2

Given the model-to-frame warp field W_t update canonical model geometry. Reconstruction into the canonical space S is represented by the sampled TSDF(truncated signed distance):

$$\nu : S \to R^2$$

Extend the projective TSDF fusion approach to operate over non-rigidly deforming scenes. Projective signed distance at the warped canonical point:

$$psdf(x_c) = [K^{-1}D_t(u_c)[u_c^T, 1]^T]_z - [x_t]_z$$

Dense Non-Rigid Surface Fusion 2/2

For each voxel x, update the TSDF to incorporate the projective SDF observed in the warped frame using TSDF fusion:

$$\nu(x)_t = \begin{cases} [v^{'}(x), w^{'}(x)]^T, psdf(dc(x)) > -\tau \\ \nu(x)_{t-1} \end{cases}$$

dc(.) transforms a discrete voxel point into the continuous TSDF domain, τ is the truncation distance.
Non-rigid fusion generalises the static reconstruction case used in KinectFusion, replacing the single (rigid) model-to-camera transform with a per voxel warp that transforms the associated space into the live frame.

Estimating the Warp-field State W_t

The objective is to update the Warp-field State W_t given a newly observed depth map D_t and the current reconstruction $\mathcal{V}$ by constructing an energy function that is minimized by the current parameters:

$$E(W_t, \mathcal{V}, D_t, \mathcal{E}) = \mathbf{Data}(W_t, \mathcal{V}, D_t) + \lambda \mathbf{Reg}(W_t, \mathcal{E})$$

Estimating the Warp-field State W_t: Dense non rigid ICP Data-term

- Transform the model $\mathcal{V}$ to a polygon mesh with point-normal pairs in the canonical frame: $\hat{\mathcal{V}}_c \equiv \{\mathcal{V}_c, \mathcal{N}_c\}$
- Transform $\hat{\mathcal{V}}_c$ to the live frame using the warp field W_t to get $\hat{\mathcal{V}}_w$
- Render $\hat{\mathcal{V}}_w$ in the live frame.
- This geometry should be close to the one in the live frame (obtained back projection of the depth image).
- Compute the error: this can be quantified by a per pixel dense model-to-frame point-plane error, which is computed under the robust Tukey penalty function data, summed over the predicted image domain augment.
- Ignoring the rendering cost, the computational complexity has an upper bound in the number of pixels in the observation

Estimating the Warp-field State W_t: Warp-field regularization

The whole problem here is: how to constrain the motion of non-observed geometry?
The assumption is made that unobserved geometry deforms in a piece-wise smooth way. To ensure this, they used deformation graph based regularization.

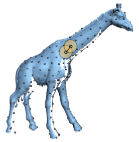


Figure 2: deformation graph example

Estimating the Warp-field State W_t: Efficient optimization

- Need an efficient solver to minimize the energy function E: used Gauss-Newton non linear optimization → forming and solving $J^T J\hat{x} = J^T e$ where $J^T J = J_d^T J_d + \lambda J_r^T J_r$
- $J_d^T J_d$ is costly (Hessian approximation): optimized by computing only block diagonal terms.
- $J_r^T J_r$ has a block arrow-head form which is efficiently factorized with a block-Cholesky decomposition.
- Improve data association by computing dense ICP (for non rigid optimization).
- For fast W computation, pre-compute, for each updated set of deformation node positions, a discretization of the k-nearest node field.

Extending the Warp-field

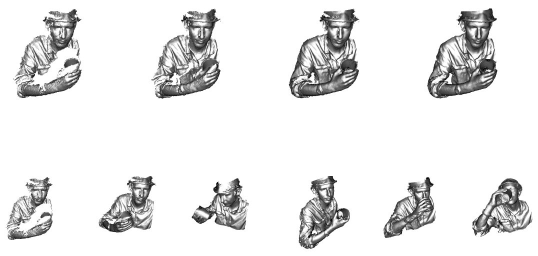
Updating the deformation graph

- After the TSDF fusion step, the polygon mesh is extracted in the canonical frame. For all vertices that are not covered by the deformation graph, sub-samble them to create new nodes that are at least ϵ distance apart.
- Initialize each new node center using the current warp
- Insert the new nodes

Updating the regularisation graph

- Induce longer range dependencies across the warp function
- Start with the deformation graph
- Sequentially decimate the graph by using the radius search based sub-sampling on the warp field nodes with an increased decimation radius.

Results



Conclusion

- Build the first real-time dense dynamic scene reconstruction system, removing the static scene assumption pervasive across real time 3D reconstruction and SLAM systems.
- Generalized TFDF fusion technique for non rigid case.
- Efficient estimation of a 6D volumetric warp field in real time
- Promising results for the future

Research

- Fusion4D: real-time performance capture of challenging scenes
- Augmented Blendshapes for Real-Time Simultaneous 3D Head Modeling and Facial Motion Capture
- Detailed, accurate, human shape estimation from clothed 3D scan sequences

Strengths

- Real time
- Non static environment
- Only need a depth camera
- No template nor prior scene model required

Weaknesses

- Problem with motion from close to open topology (opening hands)
- Model corruption due to same problems as real-time differential tracking or big motions or motion of occluded regions.
- Trade-off between stability(for highly dynamic scene) and fluid deformations.
- Memory limit due to TSDF
- Growing warp field: size and complexity of the scene increases but this also leads to more occlusion from the camera.