Co-Fusion

Introduction

- Co-Fusion is a dense SLAM system that takes a live stream of RGBD images and segments the scene into different objects while simultaneously tracking and reconstruction their 3D shape in real time.
- There are two alternative grouping strategies – motion segmentation and object instance segmentation.

Pipeline

- The system maintains two sets of object models: active models and inactive models.
- First, the system **track** the 6DOF rigid pose of each active model in current frame.
- Than, the system perform two different **segmentation** strategies.
- The last step is **fusing** the dense 3d geometry of each active model by using the newly estimated pose.

Tracking

- The cost function combines **a geometric term** based on point-to-plane ICP alignment and **a photometric color term** which is the differences in brightness between predicted color image and the current live color frame.

$$E_{icp}^{m} = \sum_{i} \left(\left(\underline{\mathbf{v^i}} - \mathbf{T_{tm}} \underline{\mathbf{v_t^i}} \right) \cdot \underline{\mathbf{n^i}} \right)^2$$

$\mathbf{v_t^i}$ is the back-ground of the vertex in current depth-map D_t

$\mathbf{v^i}$ $\mathbf{n^i}$ is the back-ground of the vertex of predicted model m from the previous frame $t-1$

$$E_{rgb}^{m} = \sum_{i} \left(\mathbf{I_t(u)} - \mathbf{I_{t-1}} \left(\pi \left(\mathbf{KT_{tm}} \pi^{-1} (\mathbf{u}, D_t) \right) \right) \right)^2$$

Motion Segmentation

- After the tracking step, the system have new estimated for the model M_t rigid transformations $\{\mathbf{T_{tm}}\}$.
- The motion segmentation problem is a labeling problem, where the labels are the M_t rigid transformations $\{\mathbf{T_{tm}}\}$.
- The cost function has two term: the unary potentials $\varphi_u(x_i)$ and the pairwise potentials $\varphi_p(x_i, x_j)$.

$$E(\underline{\mathbf{x_t}}) = \sum_i \varphi_u(x_i) + \sum_{\underline{i<j}} \varphi_p(x_i, x_j)$$

$\mathbf{x_t}$ is labeling result in the frame t

i, j are indices over the super-pixels ranging from 1 to S

Motion Segmentation

- **The unary potentials** $\varphi_u(x_i)$ are the estimated ICP alignment costs that apply the rigid transformation associated with each label as defined in E_{icp}^m. For each super-pixel with the outlier label, the cost is determined by the cost of best fitting label.
- **The pairwise potentials** $\varphi_p(x_i, x_j)$ is

$$\varphi_p(x_i, x_j) = \underline{\mu(x_i, x_j)} \sum_{m=1}^{K} w_m \underline{k_m(f_i, f_j)}$$

$\mu(x_i, x_j)$ is the classic Potts models that penalized nearby pixels taking different labels

$k_m(f_i, f_j) = \exp(-\frac{1}{2}(f_i - f_j)^{\mathrm{T}} \Lambda_m (f_i - f_j))$, where f_i f_j are the 6D feature vector and Λ_m is the inverse covariance matrix

Motion Segmentation

- After optimize the labeling, the system perform post-processing steps to merge models which have similar rigid transformations, and to spawn the new label(object) from background when the region of outliers is too large, or to put a missing label into inactive model list.

Object Instance Segmentation

- The system use the **SharpMask** to segment objects which allows to deal both with moving and static objects

Fusion

- By using the estimated pose in tracking state, the fusion stage update the surfel maps by merging the newly available RFBD frame into the existing models as the method in [8].

Conclusion

- The paper present Co-Fision, a real time RGBD SLAM system which is capable of segmenting a scene into multiple objects using motion and semantic cues, tracking and modeling them.

Other

- Github:
- https://github.com/martinruenz/co-fusion

- Website and Video:
- http://visual.cs.ucl.ac.uk/pubs/cofusion/