Multi-Object Shape Estimation and Tracking from Silhouette Cues

The Problem
- Multi-view reconstruction;
- Multiple dynamic objects in the nature environment;
- Possible static visual obstacles;
- Lighting variation, shadow, reflection, non-salient motion, background movement, color inconsistency between views, occlusion, etc.

Our Solution
- Robust probabilistic sensor fusion framework;
- m-label problem instead of two labels, m being a variable;
- Possible static visual obstacles are also recovered during the sequence;
- Automatic dynamic object appearance model initialization;
- 3D tracking, and event detection (object entering/leaving the scene)

Formulation
- Scene voxel state space: for m objects in the scene, $\mathcal{L}$ is a set of labels $\{0, 1, \ldots, m\}$;
- Observed appearance: voxel $X$ projects to $n$ views, whose colors are denoted as $c^v$, $v \in \{1, \ldots, n\}$;

If we denote $G^v_1 = (G^v_i)_{i \in \{0, 1, \ldots, m\}}$, we have the joint distribution:

$p(G^v_1, c^v_1, \ldots, c^v_n) = \prod_{v=1}^n p(G^v_1) \prod_{v=1}^n p(c^v_1 | G^v_1) \prod_{i=1}^m p(x_i | G^v_1, c^v_1, \ldots, c^v_n)$

prior terms $p(G^v_1)$ and $p(c^v_1)$

viewing line dependency terms:

$p(G^v_1 | G^v_2) = p(G^v_2) \quad \text{when } k \neq l$

image formation term:

$p(x_i | G^v_1, c^v_1, \ldots, c^v_n) = \sum_{S \in \mathcal{L}} p(x_i | S, c^v_1, \ldots, c^v_n) p(S | G^v_1, c^v_1, \ldots, c^v_n)$

where $S \in \mathcal{L}$ is the hidden silhouette state.

Bayesian Inference

$p(G^v_1, c^v_1, \ldots, c^v_n) = \frac{1}{Z} \sum_{G^v_1} p(G^v_1) \prod_{v=1}^n p(c^v_1 | G^v_1) \prod_{i=1}^m p(x_i | G^v_1, c^v_1, \ldots, c^v_n)$

where $f_k^v = \sum_{G^v_1} p(G^v_1 | g^v_k) f_k^{v+1}$ for $k < m$

and $f_m^v = \sum_{G^v_1} p(G^v_1 | g^v_k) p(x_i | g^v_k, c^v_1, \ldots, c^v_n)$

which can be simplified as:

$p(G^v_1) = \prod_{v=1}^n p(G^v_1) \prod_{k=0}^m \sum_{G^v_1} p(G^v_1 | g^v_k) p(x_i | g^v_k, c^v_1, \ldots, c^v_n)$

for $\forall v \in \{1, \ldots, n\}$.

Automatic Appearance Initialization

Object entrance criteria:
- The entrance happens only at the scene boundaries;
- $U$'s volume size is larger than a threshold;
- Subsequent updates of $U$'s track are bounded.

Learn view-based GMM model

Re-project coarse inference volume to each view

Dynamic Object’s Location Prior & Tracking

Project the probability volume on horizontal ref. plane

Search object center with a fixed window, find the most likely location

This location is used in the inference in the next time step as the location prior

Results

GMM appearance analysis. Learning per-view model allows us to bypass the tedious photometric calibration.