Motivation

• Objects exist in the scene, not in images: Images provide evidence in support of object hypotheses in 3D space.
• An object’s geometric, photometric and semantic attributes persist across multiple observations.
• Deep convolutional neural networks (CNNs) do not enforce continuity in detections across images
• Even when occluded, once seen we remain aware of objects’ presence in the scene and can predict their re-appearance.
• Objects have characteristic size (scale) and shape.
• Gravity, through inertial sensors, provides a persistent orientation reference for objects.

Object Representation

• Object attributes z (shape, pose, ID) given images up to the current time $x^t$: $p(z|x^t)$
• Condition on an attributed point cloud $s$, a minimal sufficient statistic for localization [Tsotsos et al., 2015], and sensor pose $g_t$, given images $x^t$ and inertials $u^t$ up to the current time:
  $$p(g_t, s|x^t, u^t)$$
• Marginalize object representation over viewpoint estimate from SLAM:
  $$p(z|x^t) = \int p(z|g_t, s, x^t)dP(g_t, s|x^t, u^t)$$
• Causal update of object hypotheses:
  $$p(z|g_{t+1}, s, x^{t+1}) \propto p(x^{t+1}|z, \hat{g}_t, u^t, s)p(z|g_t, s, x^t)$$

• Captures joint distribution of object shape (including scale) and identities, and geometric relations in the scene.
• Represent objects with 3D bounding boxes:

Takeaway Message

• Causal, real-time object detection and localization in the scene.
• Employs state-of-the-art real-time visual-inertial fusion/geometric mapping [Tsotsos et al., 2015] and off-the-shelf CNN (YOLO).
• Captures identities and geometric relations.
• Handles scale and occlusion.
• Future Work: Dynamic objects, topology through dense reconstruction.

References

- US Pat. Appl. 14932899