Perception 3

Wednesday, 17 March 2021 13:12

(9) Learning 3D Dynamic Scene Representation (DSR_net):

Key object properities:

1. Permanency: objects that become occulded over time; continue to exist.
2. Amodal completeness: objects have 3D occupancy; even if only partial observations are available.
3. Spatiotemporal continuity: the movement of each object is continuous in space and time.

DSR: a 3D volumetric scene representation that simultaneously discovers, tracks, reconstructs objects and predicts their dynamics.

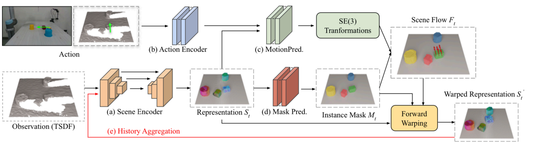

Figure 2: **DSR-Net**. DSR-Net takes in the depth observation encoded as TSDF and action as input and predicts the amodal mask of each object M_t and the voxel-wise scene flow F_t after the interaction. The scene encoder (a) outputs a representation S_t after aggregating the current observation and the history (S_t is colored by t-SNE embedding of the voxel-wise feature). S_t is then used to predict (d) amodal object instance mask M_t. In parallel, the action encoder (b) encodes the input action and the motion predictor (c) predicts object motion represented as the SE(3) transformations. The scene flow F_t is computed by combining the instance mask M_t and transformations. The warped scene representation S'_t is used as the history in the next step.

Passive perception: from passive observations (e.g. single RGB image)

Active perception: the system update the camera viewpoint for exploration and representation building.

Interactive perception: facilitated by interaction with the environment.

+: handles dynamic scenes; works with interactive perception.

(10) Learning Canonical Shape Space for Category-level 6D Object pose and size estimation (CASS):

Train a variational auto-encoder (VAE) for generating 3D point clouds in the canonical space from an RGBD image.

+: Correspondence-free approach.

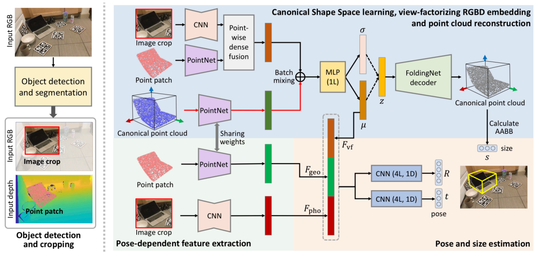

Figure 2: An overview of our network architecture. A pre-processing (left) is devised to produce image crop and point patch of the object of interest which are fed into the main network (right). The main network is composed of three modules: 1) CASS learning, pose-factorizing embedding and point cloud reconstruction (background shaded in light blue), 2) pose-dependent feature extraction (light green), and 3) pose estimation (light red). The network branch indicated with red arrows is used only in training.

In summary, the learning of the CASS and the RGBD feature embedding (denoted by F_{vf}) optimizes the following loss functions:

$$L_{\text{CASS}} = L_{\text{recon}}(X_{\text{3D}}, X^{\text{R}}_{\text{3D}}) + L_{\text{recon}}(X^{\text{R}}_{\text{rgbd}}, X^{*}_{\text{3D}}) + L_{\text{KL}}, \quad (1)$$

(input point cloud; reconstructed point cloud; reconstructed from RGBD; ground truth)

-: Cannot handle well very complex shapes due to the difficulty in reconstructing shapes with complicated geometry.

-: Cannot acheive very high precision.

-: Doesn't clos the loop in terms of utilizing the reconstructed shape geometry to guide/supervise the training of pose estimation.

(11) Learning 3D Local Descriptor for point cloud images of objects in the real world (3D_descriptor_pc):

Surface descriptor: represents the surface characteristics of an image numerically.

+: Doesn't require pre-analyzed surface information or a mesh reconstruction.

+: Directly applicable to point clouds.

+: Addresses the problem of varying densities.

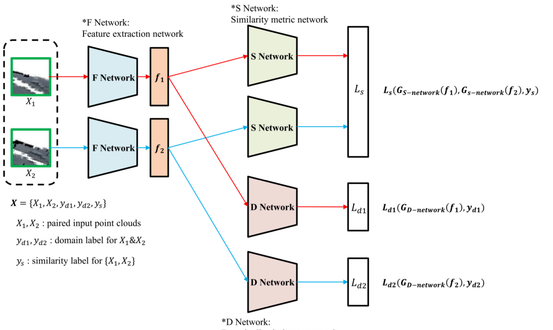

Fig. 6. The network structure used for experiments.

+: Uses Nestrov's accelerated gradient (NAG)