SceneGrok: Inferring Action Maps in 3D Environments

Manolis Savva, Angel X. Chang
Pat Hanrahan, Matthew Fisher, Matthias Nießner

Presentation: Zhenpei Yang

Outline

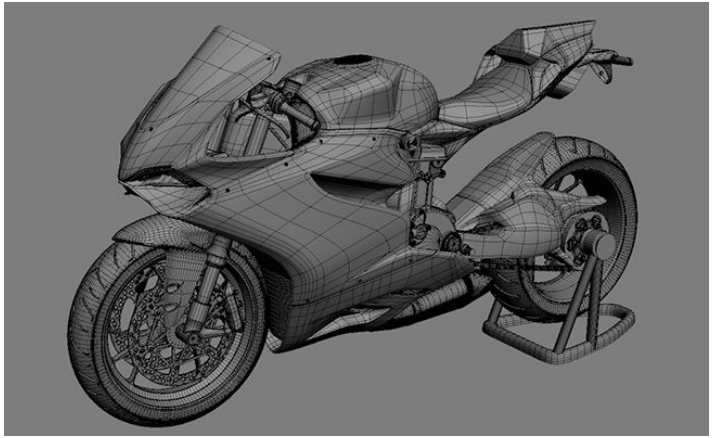
- Motivation
- Action maps: action likelihood in 3D scenes
  - Data retrieval
  - Action representation
  - Unsupervised feature learning and codebook construction
  - Scene similarity / retrieval
- Strength & Weakness

Computer Vision v.s. Computer Graphics

Motor

Computer graphics →

← Computer vision

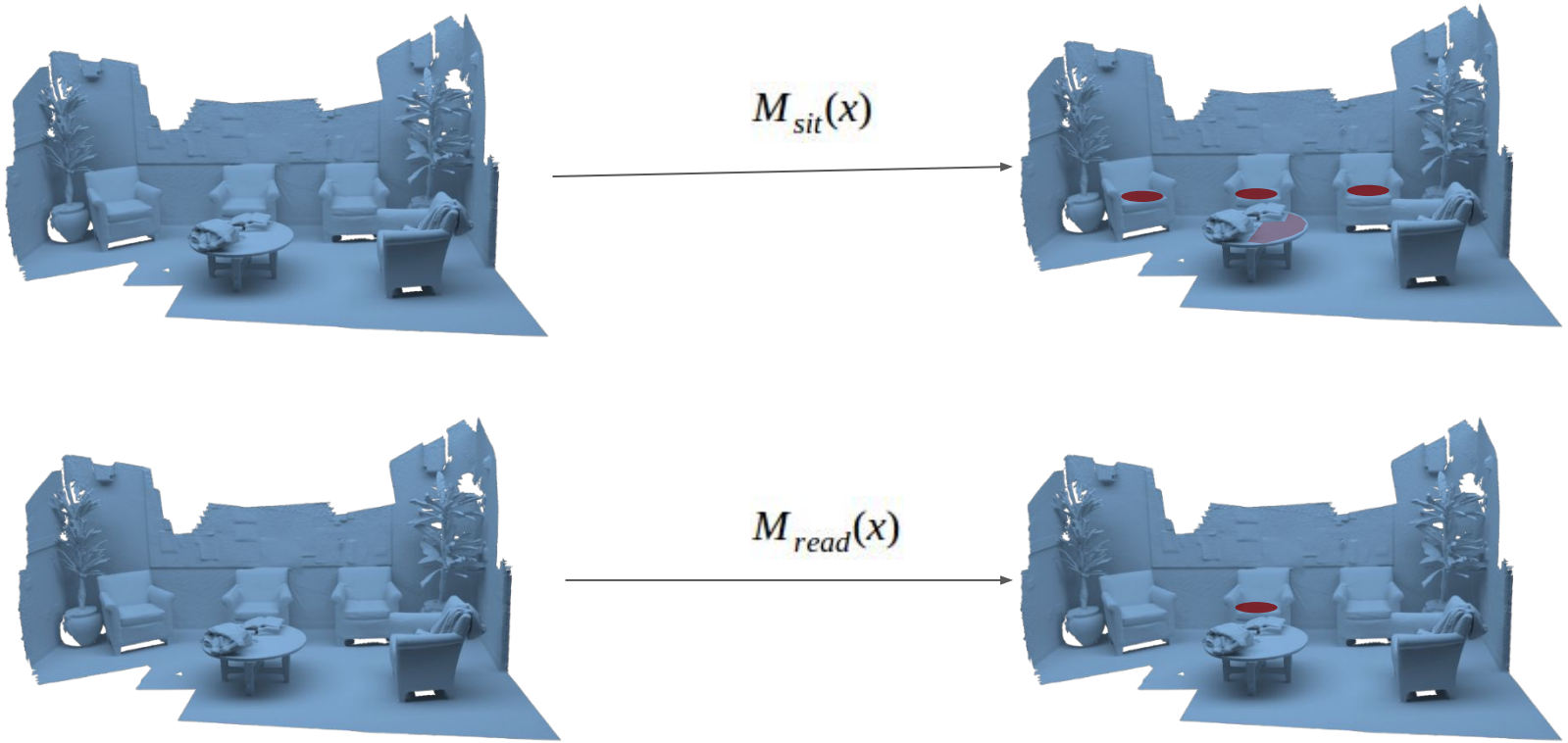


Motivation

Understand the ***semantics*** underlay 3D representations

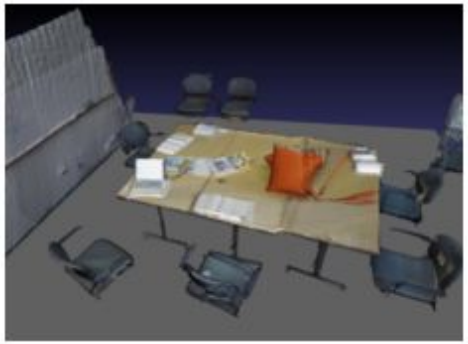
Functionality of 3D Scenes

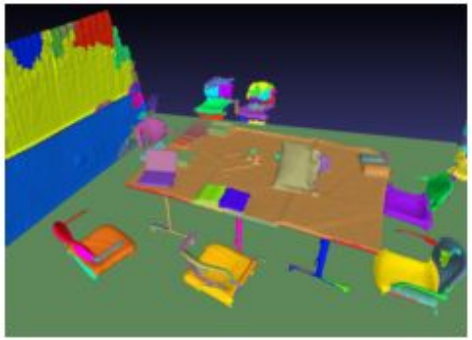
Action Map: represent the functionality of 3D Scene

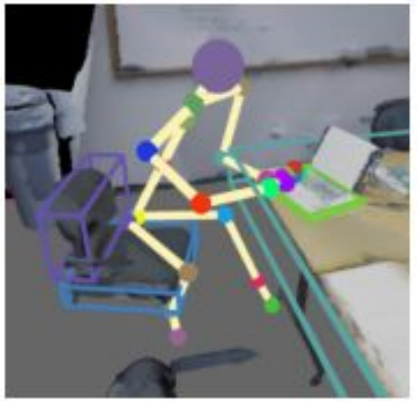


Slides credit:Manolis Savva

Data collection & Processing







Reconstructed 3D map

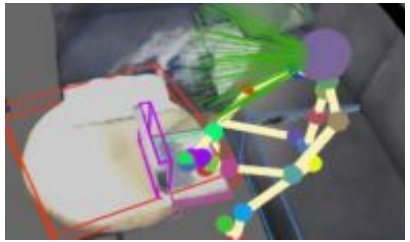
Over Segmentation

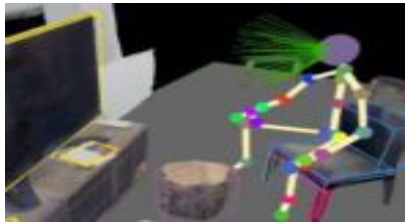
Active Segments

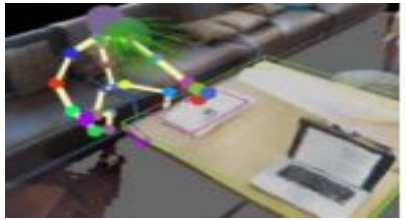
24 skeletal joints

Gaze vector

7 Actions In 14 Scenes







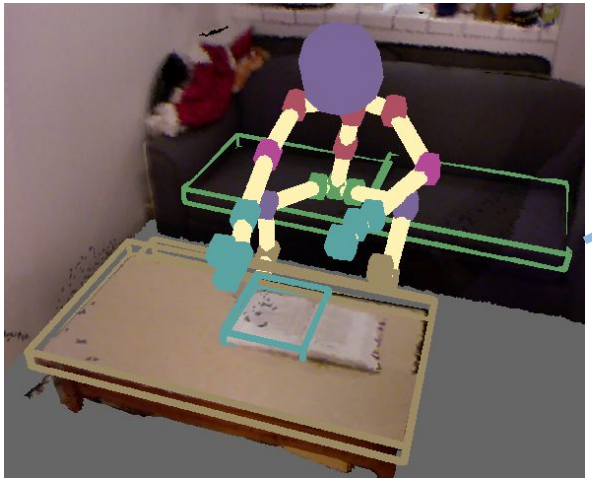
Action	Scenes	Minutes
Sit on furniture	14	54
Use a desktop computer	5	15
Read a book	10	13
Use a laptop computer	7	9
Stand on the floor	12	7
Write on a whiteboard	4	7
Watch television	4	6
Total Scenes:		14
Total Recordings:		45

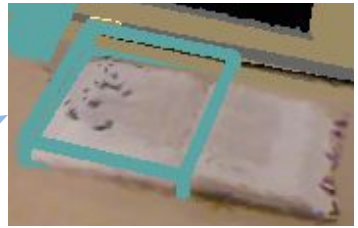
Table 1: *Summary of the dataset. For each action, we show the number of scenes in the database with at least once instance of that action and the total time spent observing this action across all recording sessions.*

How to characterize the action?

$\beta_j \in p$ → Segments activated by the jth joint of pose p

Featurization: Segment Dictionary Activation

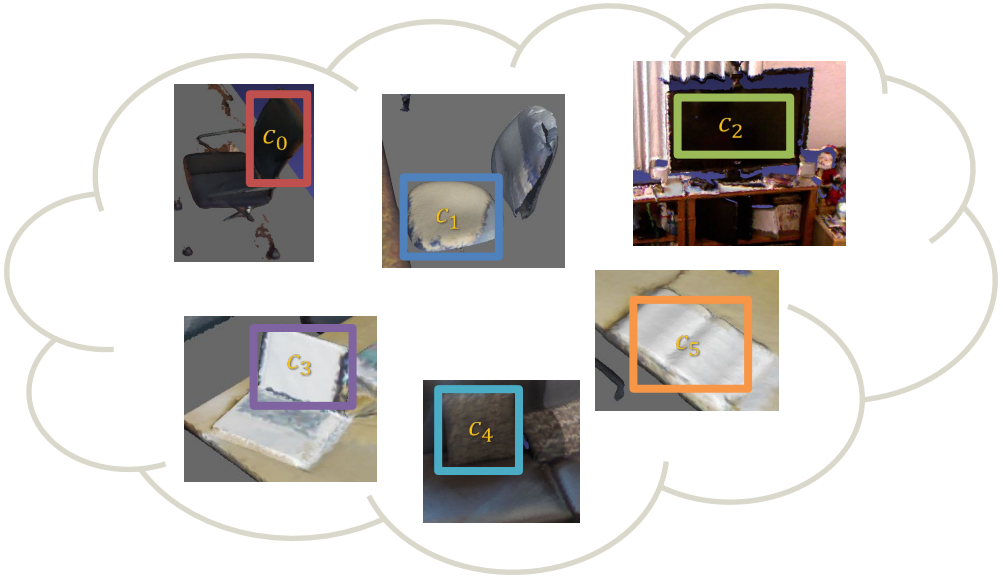




$$\begin{bmatrix} p_z \\ x_z \\ d_{xy} \\ a_{xy} \\ n_z \end{bmatrix} = \begin{bmatrix} 0.5 \\ 0.05 \\ 0.35 \\ 0.25 \\ 0.95 \end{bmatrix}$$

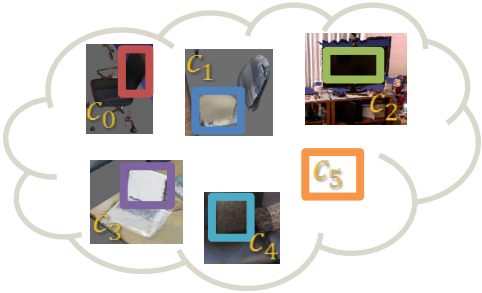
1. Vertical position of the OBB centroid above ground
2. Height of the OBB: $\max_z - \min_z$
3. Diagonal of OBB in the xy plane
4. Area of OBB in the xy plane: $\sqrt{A_{xy}(OBB)}$
5. Magnitude of the dot product of the minimum PCA vector with the world up vector.

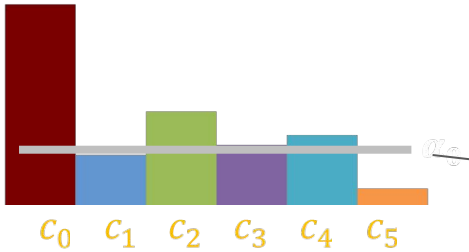
K centroids: Region Codebook



One of the centroid:

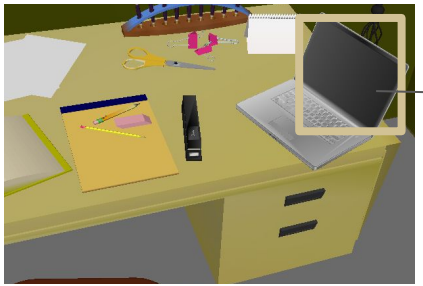
Codebook encoding



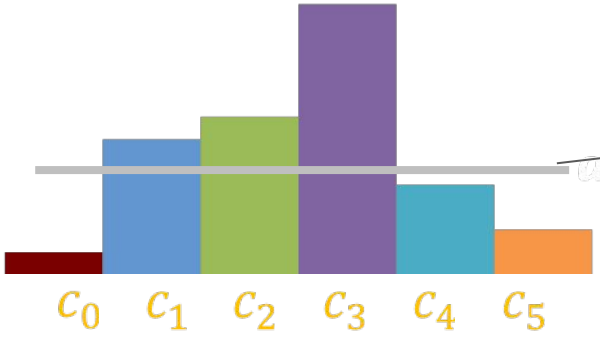


$\mathcal{D}^T s_0 \in \mathbb{R}^k$

$$\max(0, \mathcal{D}^T s - \alpha)$$



$s_1 \in \mathbb{R}^5$



$\mathcal{D}^T s_1 \in \mathbb{R}^k$

$$\max(0, \mathcal{D}^T s - \alpha)$$

Slides credit: Manolis Savva

How to characterize the action?

$\beta_j \in p$ → Segments activated by the jth joint of pose p

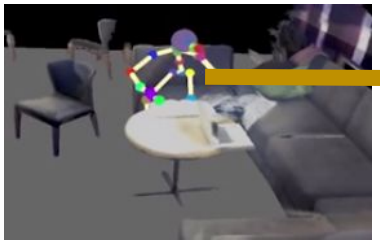
$P_a(x, y, p)$ → Probability of action a performed in pose p centered at (x,y)

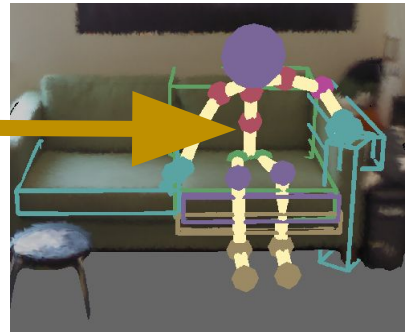
$\psi(\forall_j\{\beta_j \in p\})$ → Feature vector calculated from the activated segments of pose p

$P_a(x, y, p) \widehat{=} L(\psi(\forall_j\{\beta_j \in p\}))$ → Train a regression model L

$$M_a(x, y) = \int_{p \in H_a} P(p) P_a(x, y, p) \approx \sum_{p \in \tilde{H}_a} \frac{1}{\tilde{H}_a} P_a(x, y, p)$$ → Action map at (x,y) for action “a”

Action map prediction





Sample x,y

Sample pose

Retrieve activate segments

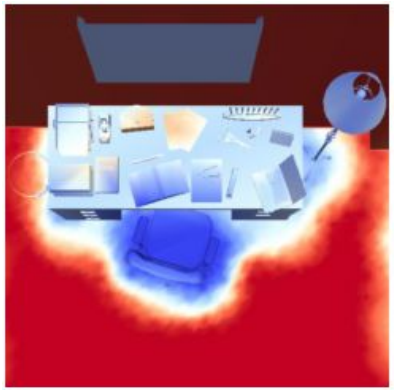
$$\begin{bmatrix} p_z \\ x_z \\ d_{xy} \\ a_{xy} \\ n_z \end{bmatrix} = \begin{bmatrix} 0.5 \\ 0.05 \\ 0.35 \\ 0.25 \\ 0.95 \end{bmatrix}$$

Extract feature

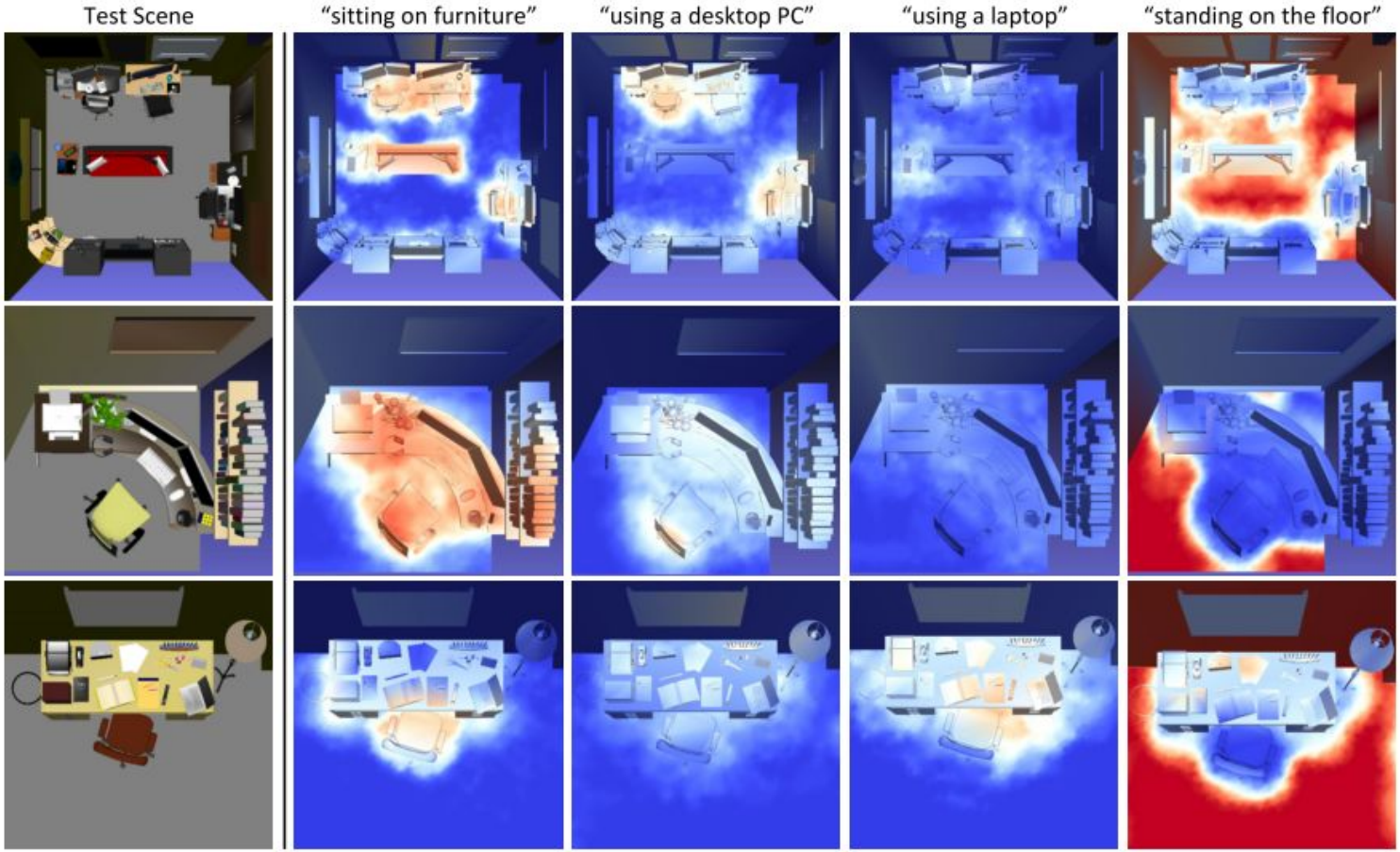
Random forest

Classification

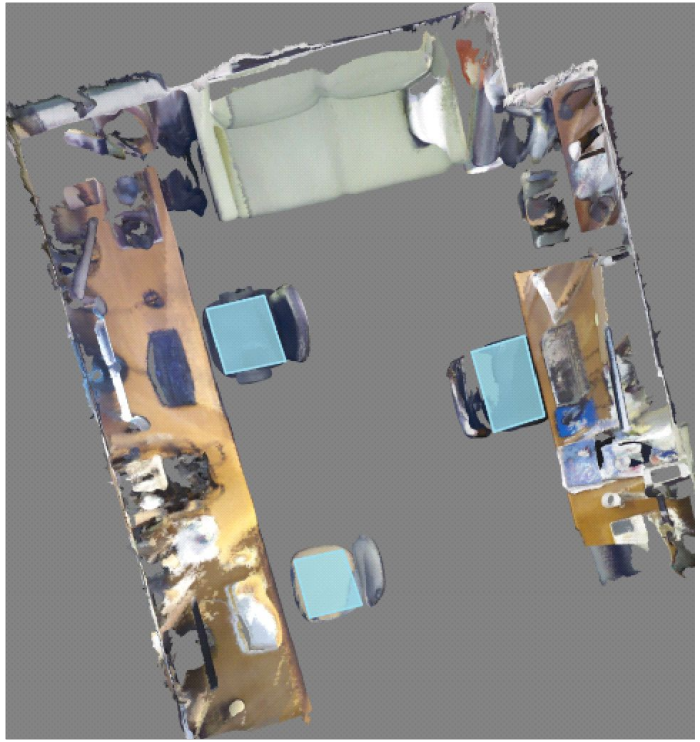
Action map



Action map prediction results:



Ground truth Labeling

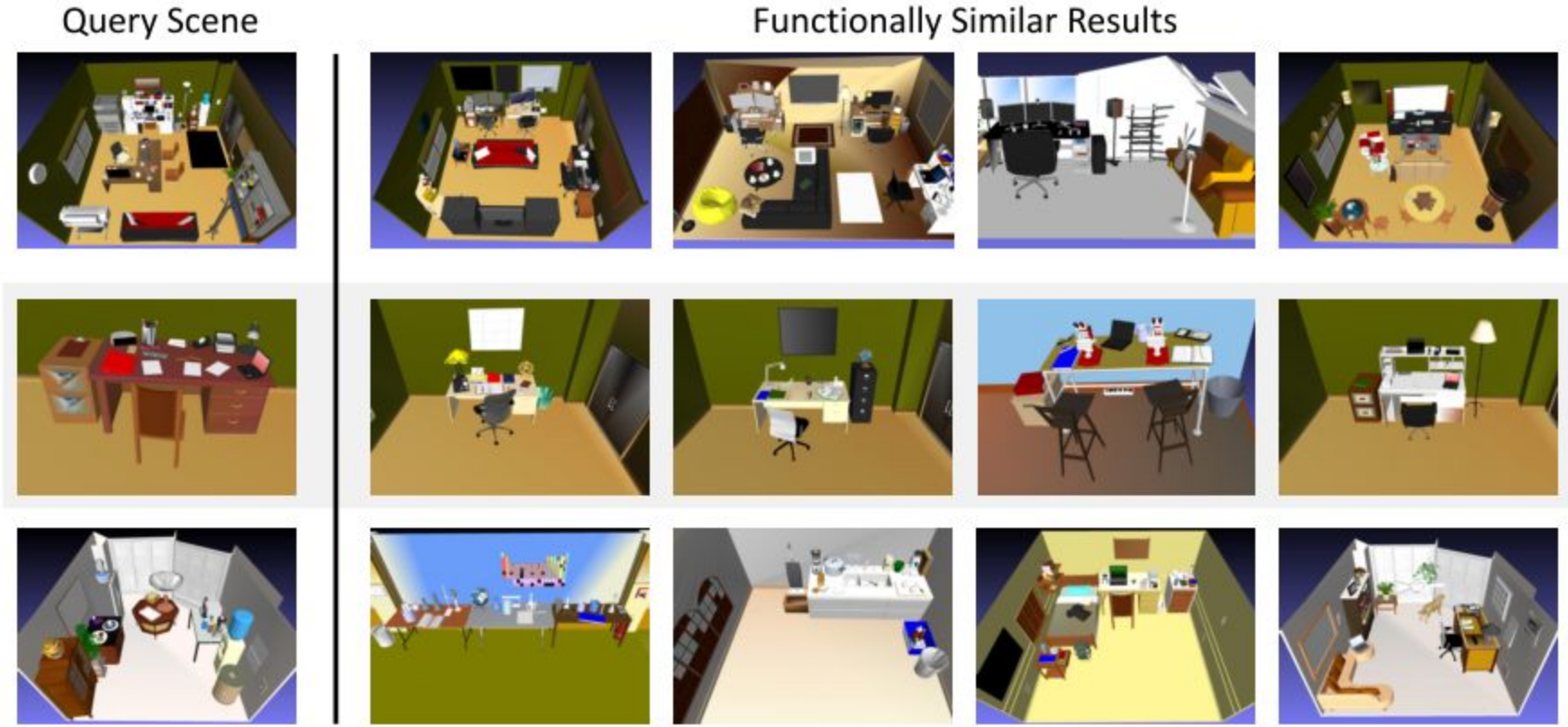


Extension: Scene retrieval through action descriptor

Scene descriptor

$$\begin{bmatrix} \int M_{sit} \\ \int M_{read} \\ \int M_{walk} \\ \int M_{sleep} \\ \vdots \end{bmatrix} \in \mathbb{R}^k$$

Functional similarity based retrieval:



Some weakness

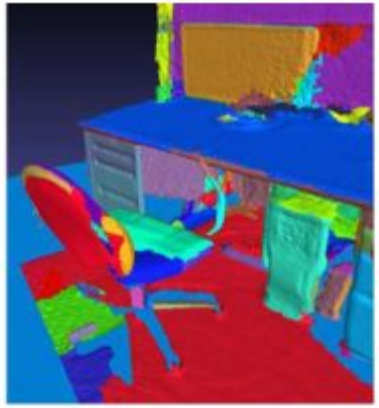
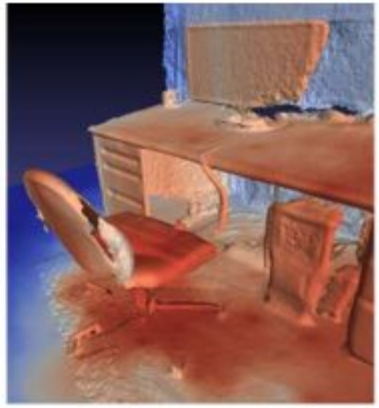
Strength

- Learn action prediction directly from real-world observation.
- Doesn't require annotated object.
- Doesn't rely on color information. User only 3D geometry.

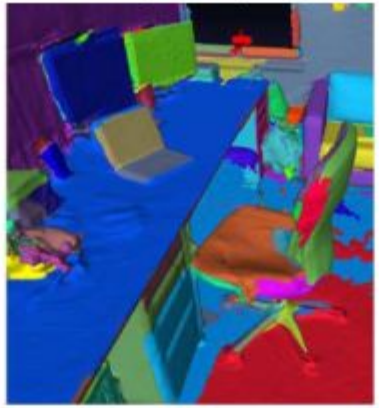
Weakness

- Cannot distinguish geometrical similar objects.
- Cannot deal with obstructions
- Ambiguity and restrictive to a small set of action.
- Data collection is expensive
- Action usually associates with objects, den

“sitting on furniture”

“using a laptop”


Possible extension

- Object-based action map
- Functional scene synthesis

Video

Thanks!