Affordances in Virtual Environments

Bhumika Kaur Matharu

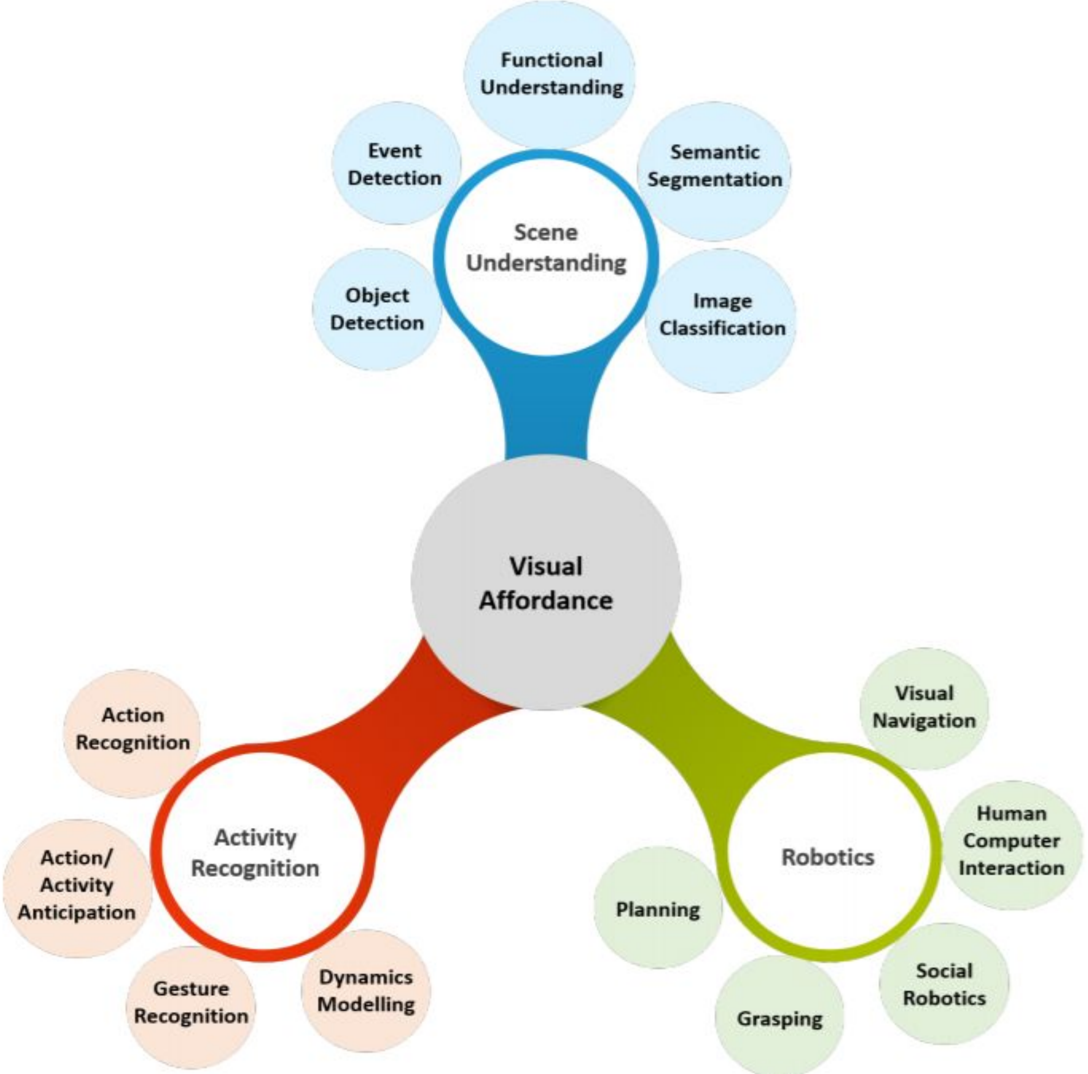
Overview of Visual Affordance

- Where is the object located in a scene?
- Object’s geometry matter, example inverted cup can’t be used for pouring water
- Should consider prior knowledge and past experience of an object, example a cup is ‘graspable’, ‘liftable’ and ‘pourable’
- A single object can take multiple affordances, example a bed is ‘sittable’ and ‘layable’ too.
- What is the action item of the object, example a stove has rotators as action item.

Research Papers

- Studied following research papers:
  - Visual Affordance and Function Understanding: A Survey (Mohammed et al.)
  - Demo2Vec: Reasoning Object Affordances from Online Videos (Kuan et al.)
  - Grounded Human - Object Interaction Hotspots from Video (Tushar et al.)

Overview of Visual Affordance



Demo2Vec Dataset - OPRA

- Online Product Review dataset for Affordance (OPRA) by collecting and labeling diverse YouTube product review videos
- Contains 11,505 demonstration clips and 2,512 object images scraped from 6 popular YouTube product review channels
- Videos include products like kitchenware objects, household appliances, consumer electronics, tools etc
- There are 7 action classes - hold, touch, rotate, push, pull, pick up, put down

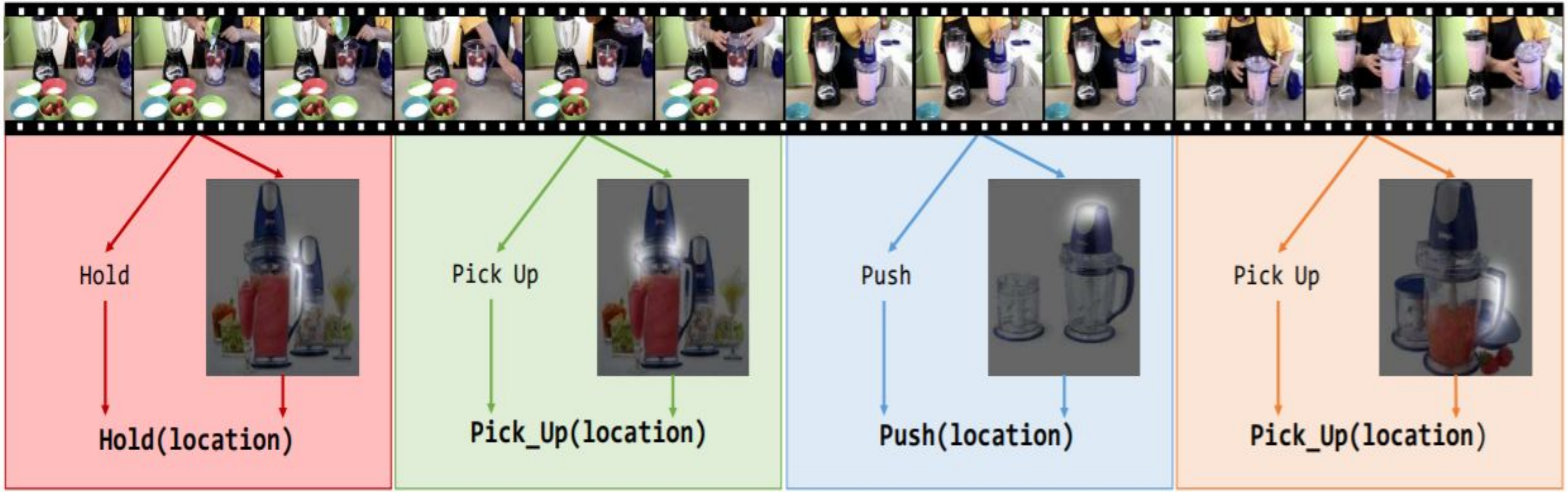
OPRA - Dataset Generation

- 1,091 full-length videos split into 2 to 15 demonstration clips
- These segmented clips contain the interaction between the user(agent) and the product
- 1 to 5 images collected for each product review video
- A total of 20,774 pairs(video+images) were generated, 16,976 for training and 3,798 for testing

Annotating Dataset

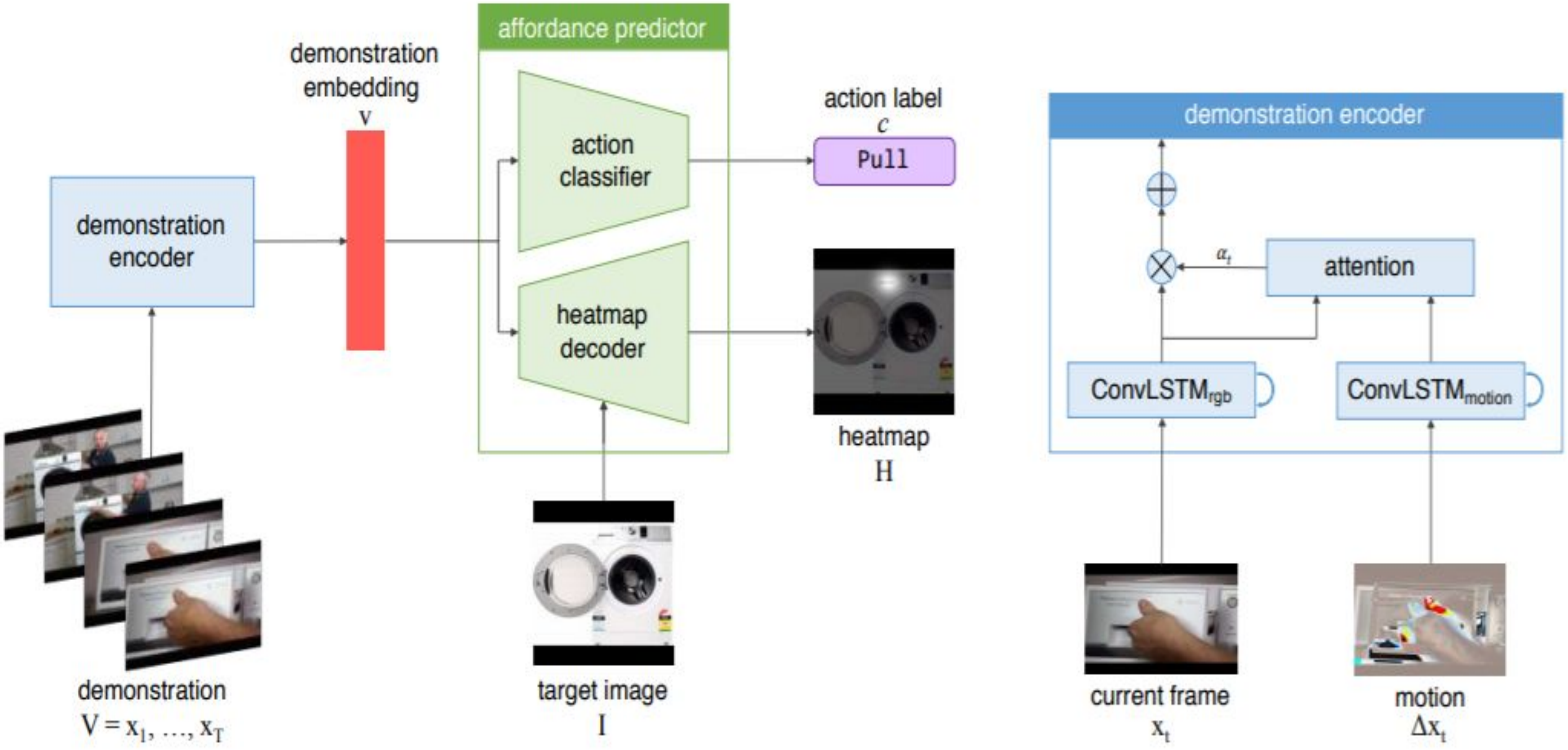
- Annotated through Amazon Mechanical Turk
- Annotator marks 10 pixels on the target image indicating the interaction region (Red points) along with action label
- Heat map is computed as a mixture of Gaussian centered at these chosen points

Example of how whole video splitted into segmented demonstration clips



Demo2Vec model

- Model composed of demonstration encoder and affordance predictor
- Extract embedded vectors from demonstration videos (Encoder)
- Predicts the interaction region and the action label on a target image of the same object (Predictor)
- Generate the heat map from the predicted interaction region
- Compare the ground truth of the heat map with the predicted heat map to calculate the loss



(a) Model overview (b) Demonstration encoder

Demo2Vec Model Overview

- Each demonstration clip consist of the interaction between the user and the object
- Camera viewpoint can change in these clips
- Target image consist of the object present in the demonstration clip
- After extracting the embedding vector from the video, affordance predictor predicts heat map and action label
- Generates predicted heat map and classify the associated action label

Implementation Details - Demonstration Encoder

- Demonstration encoder is implemented using ConvLSTM networks
- An RGB image of the object and the segmented demonstration clips are given as input to these ConvLSTM networks
- Extracts both temporal and spatial information
- Temporal soft attention mechanism added on top to aggregate the outputs
- Attention score computed by concatenating video features and image features

Implementation Details - Affordance Predictor

- Affordance Predictor consist of affordance classifier and heatmap decoder
- Affordance classifier predicts the action label on a static target image using LSTM network
- Heatmap decoder is implemented with fully convolutional layers
- Heatmap is computed by feeding concatenated embedding features extracted from demonstration encoder into transpose convolutional layers
- Softmax layer applied on top to normalize the sum of heatmap to one

Network Architecture

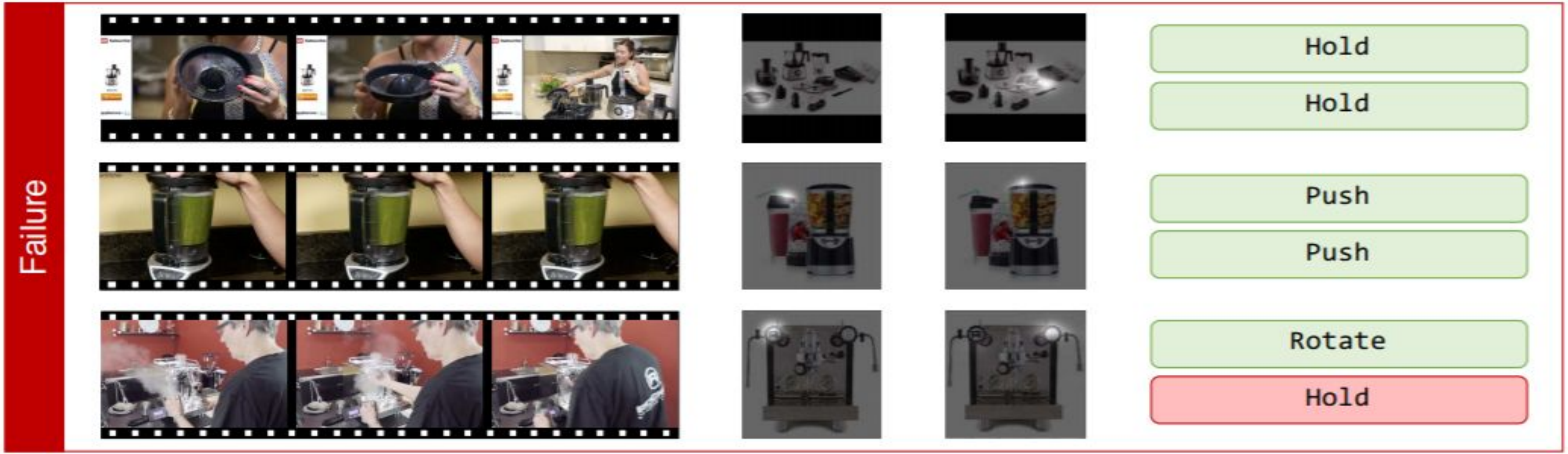
- All videos and images resized to 256X256 input
- Video subsampled to 5 FPS
- Utilized VVG16 as feature extractor trained on MS-COCO dataset
- Extracted visual representation fed into the ConvLSTMS
- Each ConvLSTMs in demonstration encoder use kernal of size 3 and stride of 1, producing recurrent of 512 channels
- Two consecutive convolutional layers applied in heatmap decoder both with 1 kernal and 1 stride
- For transposed convolutional layer, kernal of 64 and a stride of 32 applied

Training

- Took 48 hours on a single Nvidia Titan X GPU
- Adam optimizer was utilized
- Learning rate is initially set to 2 X 10**-5
- Increased the rate with decay ratio of 0.1 every 100,000 iterations
- Trained the model on 16,976 examples and tested on 3798 examples

Qualitative Results

- Able to predict heatmap and action label for a variety of scenarios and objects
- Common case of failure caused due to similar action classes
- For example, the motion of rotation often confused with holding or grasping



Conclusion

- Generated and collected real world dataset for affordance reasoning
- Number of examples in the dataset are significantly larger than existing datasets
- Model Architecture proposed achieves better performance on OPRA dataset as compared with other neural network baselines

Reference Links

- Paper 1 - https://arxiv.org/pdf/1807.06775.pdf
- Paper 2 -

  https://openaccess.thecvf.com/content_cvpr_2018/papers/Fang_Demo2Vec_

  Reasoning_Object_CVPR_2018_paper.pdf
- Paper 3 - https://arxiv.org/pdf/1812.04558.pdf