A Dataset for Developing and Benchmarking Active Vision

Phil Ammirato, Patrick Poirson, Eunbyung Park, Jana Kosecka, and Alexander C. Berg

Experiment Presentation

Presenters: Xingyi Zhou, Yajie Niu
Dataset Overview

• Dense images collection of indoor scenes

• Aligned high quality depth image.

• Bounding box and labels for object instances

• Images are connected by movement pointers
Dataset Tour

- See demo

Code provided by the authors
https://github.com/pammirato/active_vision_dataset_processing
Active Vision

• The paper used the REINFORCE algorithm for action prediction, with a reward of class scores.

• Alternative: The object score is highly related to object size, we can test simply moving forward to it, by first in-place rotating to centralize the object and then moving forward.
Active Vision - Experiment 1

- Idea: find the goal object and move towards it

- Motivation: test a simple approach on this dataset and see how it works

- Based on the intuition that when a person wants to pick up an object which is in sight, he usually catches the object with his eyes and then walk towards it.
Step 0:

Action to take: rotate to the left
Step 1:

Action to take: move forward
Step 2:

Action to take: move forward
Step 3:

Action to take: move forward
Step 4:

Action to take: move forward
Step 5:

Action to take: move forward
Step 6:

Action to take: move forward
Step 7:

Action to take: move forward
Can’t move forward anymore.
Problem: can't go around the obstacle on the way

Action to take: rotate to the left
Problem: unexpected position change when rotating

Action to take: rotate to the left
Problem: unexpected position change when rotating

Step 1:

Action to take: rotate to the left
Problem: unexpected position change when rotating

Step 2:

Action to take: rotate to the left
Problem: unexpected position change when rotating

Step 3:

Action to take: rotate to the left

A sudden change of position!
Problem: unexpected position change when rotating

Step 4:

Action to take: rotate to the left
Problem: unexpected position change when rotating

Step 5:
# Alternative 1 - Results

## Results

<table>
<thead>
<tr>
<th>Method</th>
<th>Split 1</th>
</tr>
</thead>
<tbody>
<tr>
<td>REINFORCE</td>
<td>0.45</td>
</tr>
<tr>
<td>Alternative 1</td>
<td>0.330</td>
</tr>
<tr>
<td>Random</td>
<td>0.208</td>
</tr>
</tbody>
</table>

## Drawbacks

- Can’t bypass the obstacle on the way
- Position change due to the dataset
- ‘Fine-tuning’ at the end to get a higher accuracy score
Active Vision - Supervised

- Alternative 2: Since we have all the object score information in training, we can apply supervised learning guided by the ground truth best movement.
Active Vision - Supervised

- Training data generation
  - Each frame is a tuple of (image, bbox, target_object_score)
  - Assign one of the six directions or a stop sign as classification target. Score is discarded in training.

(Image, box, score = 0.4)

Assigned action: rotate clockwise

(score = 0.35)  (score = 0.8)  (score = 0.9)
Supervised - Framework

Stacked RGB + Object Mask

- input is a 4 channel RGB+Mask tensor
- The convolutional weight of the first 3 channel is copied from pretrained resnet
- initialize the conv weight of Mask channel with zero, so in the initial stage the resnet performs exactly the same as 3-channel version.

“What happens if…” Learning to Predict the Effect of Forces in Images. Mottaghi, R., Rastegari, M., Gupta, A., & Farhadi, A. ECCV16
## Supervised - Results

<table>
<thead>
<tr>
<th>Number of Moves</th>
<th>5</th>
<th>20</th>
</tr>
</thead>
<tbody>
<tr>
<td>Method</td>
<td>Split 1</td>
<td></td>
</tr>
<tr>
<td>REINFORCE</td>
<td>0.45</td>
<td>0.51</td>
</tr>
<tr>
<td>Greedy</td>
<td>0.330</td>
<td>0.394</td>
</tr>
<tr>
<td>Random</td>
<td>0.208</td>
<td>0.251</td>
</tr>
<tr>
<td>Supervised</td>
<td>0.252</td>
<td>0.304</td>
</tr>
</tbody>
</table>

Problem: The robot is easy to get stuck in a cycle or a deadend.

Code modified from the authors, by Xingyi Zhou
https://github.com/xingyizhou/deep_active_vision
Supervised - Demos

<table>
<thead>
<tr>
<th>Episode</th>
<th>object_id</th>
<th>image_id</th>
<th>correct</th>
<th>score</th>
<th>bb</th>
</tr>
</thead>
<tbody>
<tr>
<td>1041</td>
<td>29</td>
<td>816</td>
<td>0</td>
<td>0.0168</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>805</td>
<td>0</td>
<td>0.0429</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>793</td>
<td>1</td>
<td>0.6805</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>804</td>
<td>0</td>
<td>0.2034</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>793</td>
<td>1</td>
<td>0.6805</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>804</td>
<td>0</td>
<td>0.2034</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>793</td>
<td>1</td>
<td>0.6805</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>804</td>
<td>0</td>
<td>0.2034</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>793</td>
<td>1</td>
<td>0.6805</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>804</td>
<td>0</td>
<td>0.2034</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>793</td>
<td>1</td>
<td>0.6805</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>804</td>
<td>0</td>
<td>0.2034</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>793</td>
<td>1</td>
<td>0.6805</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>804</td>
<td>0</td>
<td>0.2034</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>793</td>
<td>1</td>
<td>0.6805</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>804</td>
<td>0</td>
<td>0.2034</td>
<td></td>
</tr>
</tbody>
</table>
Supervised - Demos
Supervised - Demos
Supervised - Demos
Supervised - Demos
Supervised - Demos

```plaintext
Episodes 1054: object_id 20:
image_id: 830, correct: 0, score: 0.0068
action: 6, image_id: 830, correct: 0, score: 0.0068, bb(75,118,78,131)
action: 5, image_id: 829, correct: 0, score: 0.0000, bb(1,1,2,2)
action: 2, image_id: 817, correct: 0, score: 0.0000, bb(1,1,2,2)
action: 6, image_id: 818, correct: 0, score: 0.0000, bb(1,1,2,2)
action: 3, image_id: 773, correct: 0, score: 0.0000, bb(1,1,2,2)
action: 3, image_id: 701, correct: 0, score: 0.0000, bb(1,1,2,2)
action: 3, image_id: 653, correct: 0, score: 0.0000, bb(1,1,2,2)
action: 5, image_id: 652, correct: 0, score: 0.0000, bb(1,1,2,2)
action: 4, image_id: 712, correct: 0, score: 0.0000, bb(1,1,2,2)
action: 6, image_id: 713, correct: 0, score: 0.0000, bb(1,1,2,2)
action: 3, image_id: 665, correct: 0, score: 0.0000, bb(1,1,2,2)
action: 1, image_id: 701, correct: 0, score: 0.0000, bb(1,1,2,2)
action: 3, image_id: 653, correct: 0, score: 0.0000, bb(1,1,2,2)
action: 5, image_id: 652, correct: 0, score: 0.0000, bb(1,1,2,2)
action: 4, image_id: 712, correct: 0, score: 0.0000, bb(1,1,2,2)
action: 6, image_id: 713, correct: 0, score: 0.0000, bb(1,1,2,2)
action: 3, image_id: 665, correct: 0, score: 0.0000, bb(1,1,2,2)
action: 1, image_id: 701, correct: 0, score: 0.0000, bb(1,1,2,2)
action: 3, image_id: 653, correct: 0, score: 0.0000, bb(1,1,2,2)
action: 5, image_id: 652, correct: 0, score: 0.0000, bb(1,1,2,2)
```
Supervised - Demos
Supervised - Demos
Supervised - Demos
Supervised - Demos
Supervised - Demos
Conclusion

• Dataset tour

• Experiment 1: moving towards the goal object through a straight line

• Experiment 2: supervised learning given the ground truth best action.

• Active vision is a challenging task and this dataset serves as a useful benchmark for this task.
Thank you!