Presentation (paper review)
Learning Image Representations Tied to Ego-motion
Jayaraman and Grauman. ICCV 2015

Hilgad Montelo
Outline

- The "Kitten Carousel" Experiment
- Problem
- Objective
- Main Idea
- Related Work
- Approach
- Experiments and Results
- Conclusions
The "Kitten Carousel" Experiment (Held & Hein, 1963)

Key to perceptual development:
self-generated motion + visual feedback

[Slide credit: Dinesh Jayaraman]
Problem

• Today’s visual recognition algorithms learn from “disembodied” bag of labeled snapshots.
Objective

• Provide visual recognition algorithm that learns in the context of **acting** and **moving** in the world.
Main Idea

- Associate Ego-Motion and vision by teaching computer vision system the connection:
- "how I move" ↔ "how my visual surroundings change"
Ego-motion ↔ vision: view prediction

After moving:

[Slide credit: Dinesh Jayaraman]
Ego-motion ↔ vision for recognition

• Learning this connection requires:
  ➢ Depth, 3D geometry
  ➢ Semantics
  ➢ Context

• Can be learned without manual labels!

**Approach:** unsupervised feature learning using egocentric video + motor signals
Related Works

Integrating vision and motion
Agrawal, Carreira, Malik, “Learning to see by moving”, ICCV 2015
Watter, Springenberg, Boedecker, Riedmiller, “Embed to control...”, NIPS 2015
Levine, Finn, Darrell, Abbeel, “... visuomotor policies”, arXiv 2015
Konda, Memisevic, “Learning visual odometry ...”, VISAPP 2015

Visual prediction
Doersch, Gupta, Efros, “... context prediction”, ICCV 2015
Kulkarni, Whitney, Kohli, Tenenbaum, “... inverse graphics ...”, NIPS 2015

Video for unsupervised image features
Wang, Gupta, “Unsupervised learning of visual ...”, ICCV 2015
Approach

Ego-motion equivariance

**Invariant features:** unresponsive to some classes of transformations

\[ z(gx) \approx z(x) \]

**Equivariant features:** *predictably* responsive to some classes of transformations, through simple mappings (e.g., linear)

\[ z(gx) \approx \hat{M}_g z(x) \]

Invariance *discards* information; equivariance *organizes* it.
Approach

**Training data**
Unlabeled video + motor signals

**Equivariant embedding**
organized by ego-motions

Pairs of frames related by similar ego-motion should be related by same feature transformation

Source: “Learning image representations equivariant to ego motion”, Jayaraman and Grauman, ICCV 2015
Approach

1. Extract training frame pairs from video
2. Learn ego-motion-equivariant image features
3. Train on target recognition task in parallel
Training frame pair mining

Discovery of ego-motion clusters

- $g =$left turn
- $g =$forward
- $g =$right turn

[yaw change] [forward distance]

[Slide credit: Dinesh Jayaraman]
Ego-motion equivariant feature learning

**Given:**

- $x_i$,
- $g x_i$

**Desired:** for all motions $g$ and all images $x$,

$$z_\theta(gx) \approx M_g z_\theta(x)$$

**Unsupervised training**

- $z_\theta(x_i)$
- $M_g$
- $\| M_g z_\theta(x_i) - z_\theta(g x_i) \|_2$

**Supervised training**

- $x_k$
- Class $y_k$
- $z_\theta(x_i)$

$\theta, M_g$ and $W$ jointly trained

[Slide credit: Dinesh Jayaraman]
Experiments

• Validation using 3 public datasets: NORB, KITTI, SUN.
• Comparison with different methods: CLSNET, TEMPORAL, DRLIM.
Results: Recognition

Learn from **unlabeled car video** (KITTI)

Exploit features for **static scene classification**
(SUN, 397 classes)

Geiger et al, IJRR ’13

Xiao et al, CVPR ’10
Results: Recognition

Do ego-motion equivariant features improve recognition?

Up to 30% accuracy increase over state of the art!

*Hadsell et al., Dimensionality Reduction by Learning an Invariant Mapping, CVPR'06

**Mobahi et al., Deep Learning from Temporal Coherence in Video, ICML’09
Results: Active recognition

- Leverage proposed equivariant embedding to select next best view for object recognition.

![Cup and frying pan images with labels and accuracy graph]

NORB data

Accuracy (%)

Random
DRLim [Hadsell et al.]
Temporal [Mobahi et al.]
Ours

[Slide credit: Dinesh Jayaraman]
Conclusion and Future Work

• The paper provided a new *embodied visual feature learning* paradigm.

• The *Ego-motion equivariance* boosts performance across multiple challenging recognition tasks.
Questions

• Why KITTI training and not some other domain based training?
• Why does incorporating DRLIM improve EQUIV? Still Temporal coherence properties left to be learned?
• Is it meaningful to compare EQUIV or EQUIV + DRLIM with the other cases with respect to equivariance error?
Thank You