Learning image representations tied to ego-motion

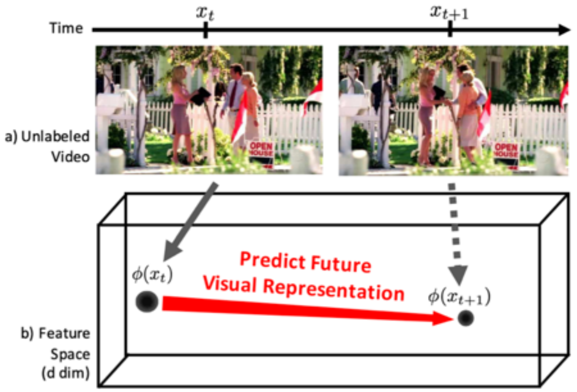
Dinesh Jayaraman
The University of Texas at Austin
dineshj@cs.utexas.edu

Kristen Grauman
The University of Texas at Austin
grauman@cs.utexas.edu

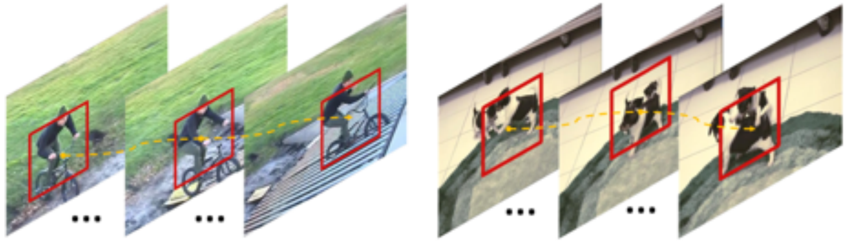
ICCV 2015

Motivation

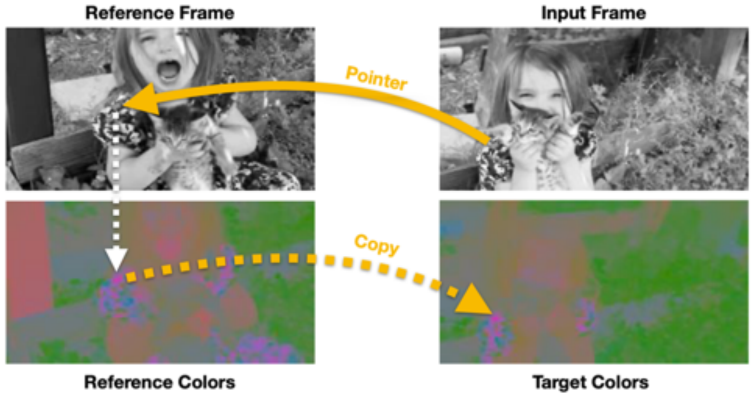
- Self-supervised learning approaches for videos
  - Future visual representation prediction
  - Tracking objects
  - Colorization
  - More…?



Vondrick et. al, CVPR 2016



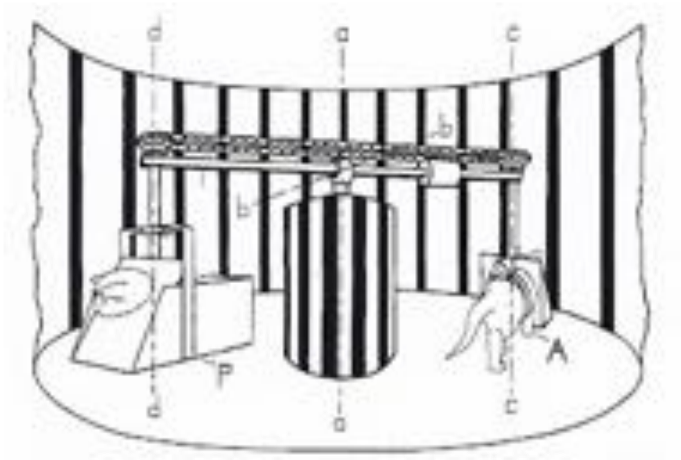
Wang et. al, ICCV 2015



Vondrick et. al, ECCV 2018

Motivation

- Kitten carousel experiment [Held, R. and Hein A. (1963)]



- learn representations that exploit the parallel signals of ego-motion and pixels

Methods

- High level ideas
  - Mining ego-motion patterns from video
  - Learning the transformation between image (feature) pairs
  - Regularizing (incorporating) a recognition task

Methods – Mining Ego-motion Patterns

- Organize training sample pairs into a discrete set of ego-motion patterns

- Apply k-means to obtain G clusters, with $p \in \{1, .., G\}$ denotes motion pattern ID

- Input & Label: $\langle (x_i, x_j), p_{ij} \rangle$

Methods – Ego-motion Equivariance

- $\forall \boldsymbol{x} \in \mathcal{X} : \mathbf{z}_{\boldsymbol{\theta}}(g\boldsymbol{x}) \approx M_g \mathbf{z}_{\boldsymbol{\theta}}(\boldsymbol{x}).$

- z_θ: feature space, M_g: equivariance map

- M_g represents the affine transformation in the feature space that corresponds to transformation g in the pixel space

- "the learned feature space will *not* be limited to preserving equivariance for pairs originating from the same ego-motions" -> why?

Methods – Ego-motion Equivariance

- $\forall \boldsymbol{x} \in \mathcal{X} : \mathbf{z}_{\boldsymbol{\theta}}(g\boldsymbol{x}) \approx M_g \mathbf{z}_{\boldsymbol{\theta}}(\boldsymbol{x}).$
- z_θ: feature space, M_g: equivariance map
- M_g represents the affine transformation in the feature space that corresponds to transformation g in the pixel space
- “the learned feature space will *not* be limited to preserving equivariance for pairs originating from the same ego-motions” -> why?

$$\mathbf{z}(d\boldsymbol{x}) = \mathbf{z}((r \circ u)\boldsymbol{x}) = M_r \mathbf{z}(u\boldsymbol{x}) = M_r M_u \mathbf{z}(\boldsymbol{x})$$

Methods – Learning Objective

- $$(\boldsymbol{\theta}^*, \mathcal{M}^*) = \underset{\boldsymbol{\theta}, \mathcal{M}}{\arg\min} \sum_{g} \sum_{\{(i,j):p_{ij}=g\}} d\left(M_g \mathbf{z}_{\boldsymbol{\theta}}(\boldsymbol{x}_i), \mathbf{z}_{\boldsymbol{\theta}}(\boldsymbol{x}_j)\right)$$
  - Shall not work, why?

Methods – Learning Objective

- $$(\boldsymbol{\theta}^*, \mathcal{M}^*) = \arg\min_{\boldsymbol{\theta}, \mathcal{M}} \sum_{g} \sum_{\{(i,j): p_{ij}=g\}} d\left(M_g \mathbf{z}_{\boldsymbol{\theta}}(\boldsymbol{x}_i), \mathbf{z}_{\boldsymbol{\theta}}(\boldsymbol{x}_j)\right)$$
  - Shall not work, why?

- $$(\boldsymbol{\theta}^*, \mathcal{M}^*) = \arg\min_{\boldsymbol{\theta}, \mathcal{M}} \sum_{g,i,j} d_g\left(M_g \mathbf{z}_{\boldsymbol{\theta}}(\boldsymbol{x}_i), \mathbf{z}_{\boldsymbol{\theta}}(\boldsymbol{x}_j), p_{ij}\right)$$

$$\begin{aligned} d_g(\boldsymbol{a}, \boldsymbol{b}, c) = \mathbb{1}(c = g) d(\boldsymbol{a}, \boldsymbol{b}) + \\ \mathbb{1}(c \neq g) \max(\delta - d(\boldsymbol{a}, \boldsymbol{b}), 0), \end{aligned}$$

Methods – Regularizing in a Recognition Task

- Suppose in addition to ego-annotated pairs $\mathcal{U}$, also given a small set of class-labeled static images $\mathcal{L}$

- $(\boldsymbol{\theta}^*, W^*, \mathcal{M}^*) = \underset{\boldsymbol{\theta}, W, \mathcal{M}}{\arg\min}\, L_c(\boldsymbol{\theta}, W, \mathcal{L}) + \lambda L_e(\boldsymbol{\theta}, \mathcal{M}, \mathcal{U}),$

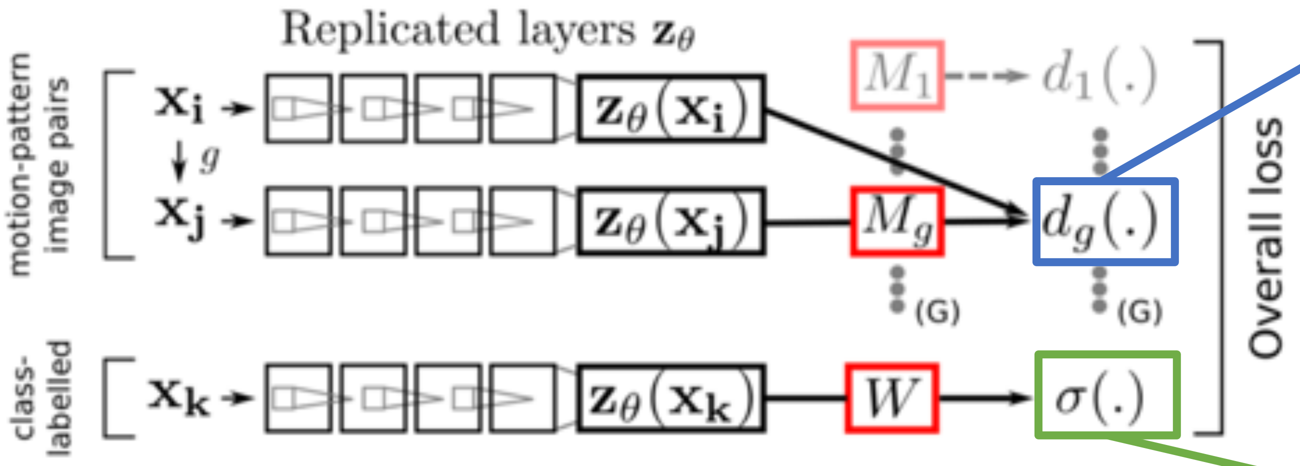
  - $L_c(W, \mathcal{L}) = -\frac{1}{N_l} \sum_{i=1}^{N_l} \log(\sigma_{c_k}(W \mathbf{z}_{\boldsymbol{\theta}}(\boldsymbol{x}_i))$

  - W is classifier weights, σ is the softmax probability

Methods – Learning the z_θ

•

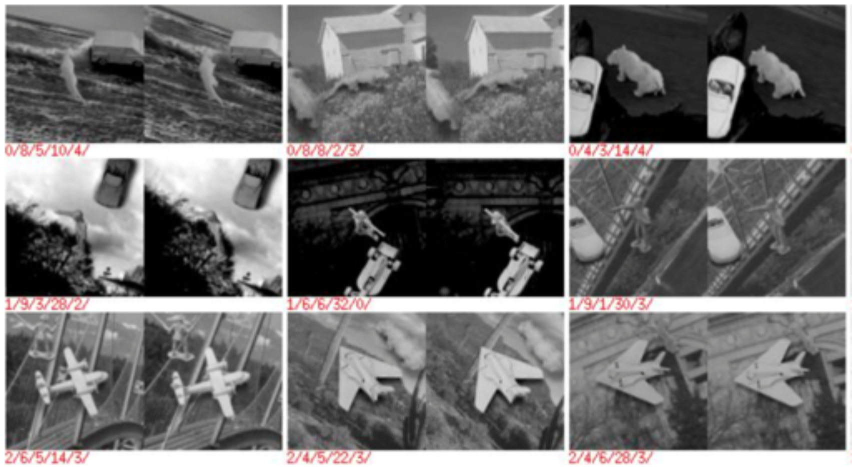
$$(\boldsymbol{\theta}^*, \mathcal{M}^*) = \underset{\boldsymbol{\theta}, \mathcal{M}}{\arg\min} \sum_{g,i,j} d_g\left(M_g \mathbf{z}_{\boldsymbol{\theta}}(\boldsymbol{x}_i), \mathbf{z}_{\boldsymbol{\theta}}(\boldsymbol{x}_j), p_{ij}\right)$$



$$\underset{\boldsymbol{\theta}, W, \mathcal{M}}{\arg\min} L_c(\boldsymbol{\theta}, W, \mathcal{L})$$

Experiments

- Target tasks
  - Recognition
  - Next-best view
    - Given one view of a object, tell a robot how to move to help recognize the object
- Datasets
  - Unsupervised learning
    - NORB
    - KITTI
  - Supervised learning
    - NORB
    - KITTI
    - SUN



Experiments

- Baselines
  - CLSNET - trained on only supervised samples
  - TEMPORAL - temporal coherence $$\boldsymbol{\theta}^* = \arg\min_{\boldsymbol{\theta}} \sum_{i,j} d_1(\mathbf{z}_{\boldsymbol{\theta}}(\boldsymbol{x}_i), \mathbf{z}_{\boldsymbol{\theta}}(\boldsymbol{x}_j), \mathbb{1}(|t_i - t_j| \leq T))$$
  - DRLIM - TEMOPRAL but with $\ell 2$ distance

Quantitative Results

Tasks→	Equivariance error		Recognition accuracy %				Next-best view
Datasets→	NORB		NORB-NORB	KITTI-KITTI	KITTI-SUN	KITTI-SUN	NORB
Methods↓	atomic	composite	[25 cls]	[4 cls]	[397 cls]	[397 cls, top-10]	1-view→ 2-view
random	1.0000	1.0000	4.00	25.00	0.25	2.52	4.00 → 4.00
CLSNET	0.9239	0.9145	25.11±0.72	41.81±0.38	0.70±0.12	6.10±0.67	-
TEMPORAL [21]	0.7587	0.8119	35.47±0.51	45.12±1.21	1.21±0.14	8.24±0.25	29.60→ 31.90
DRLIM [9]	0.6404	0.7263	36.60±0.41	47.04±0.50	1.02±0.12	6.78±0.32	14.89→ 17.95
EQUIV	**0.6082**	**0.6982**	**38.48±0.89**	**50.64±0.88**	**1.31±0.07**	**8.59±0.16**	**38.52→43.86**
EQUIV+DRLIM	**0.5814**	**0.6492**	**40.78±0.60**	**50.84±0.43**	**1.58±0.17**	**9.57±0.32**	**38.46→43.18**

Table 1. (Left) Average equivariance error (Eq (9)) on NORB for ego-motions like those in the training set (atomic) and novel ego-motions (composite). (Center) Recognition result for 3 datasets (mean ± standard error) of accuracy % over 5 repetitions. (Right) Next-best view selection accuracy %. Our method EQUIV (and augmented with slowness in EQUIV+DRLIM) clearly outperforms all baselines.

Equivariance Error: $\rho_g = E\left[\|\mathbf{z}_{\boldsymbol{\theta}}(\boldsymbol{x}) - M_g^{'}\mathbf{z}_{\boldsymbol{\theta}}(g\boldsymbol{x})\|_2 / \|\mathbf{z}_{\boldsymbol{\theta}}(\boldsymbol{x}) - \mathbf{z}_{\boldsymbol{\theta}}(g\boldsymbol{x})\|_2\right]$

Quantitative Results

Tasks→ Datasets→ Methods↓	Equivariance error NORB atomic	composite	Recognition accuracy % NORB-NORB [25 cls]	KITTI-KITTI [4 cls]	KITTI-SUN [397 cls]	KITTI-SUN [397 cls, top-10]	Next-best view NORB 1-view→ 2-view
random	1.0000	1.0000	4.00	25.00	0.25	2.52	4.00 → 4.00
CLSNET	0.9239	0.9145	25.11±0.72	41.81±0.38	0.70±0.12	6.10±0.67	-
TEMPORAL [21]	0.7587	0.8119	35.47±0.51	45.12±1.21	1.21±0.14	8.24±0.25	29.60→ 31.90
DRLIM [9]	0.6404	0.7263	36.60±0.41	47.04±0.50	1.02±0.12	6.78±0.32	14.89→ 17.95
EQUIV	**0.6082**	**0.6982**	**38.48±0.89**	**50.64±0.88**	**1.31±0.07**	**8.59±0.16**	**38.52→43.86**
EQUIV+DRLIM	**0.5814**	**0.6492**	**40.78±0.60**	**50.84±0.43**	**1.58±0.17**	**9.57±0.32**	**38.46→43.18**

Table 1. (Left) Average equivariance error (Eq (9)) on NORB for ego-motions like those in the training set (atomic) and novel ego-motions (composite). (Center) Recognition result for 3 datasets (mean ± standard error) of accuracy % over 5 repetitions. (Right) Next-best view selection accuracy %. Our method EQUIV (and augmented with slowness in EQUIV+DRLIM) clearly outperforms all baselines.

Quantitative Results

Tasks→	Equivariance error		Recognition accuracy %				Next-best view
Datasets→	NORB		NORB-NORB	KITTI-KITTI	KITTI-SUN	KITTI-SUN	NORB
Methods↓	atomic	composite	[25 cls]	[4 cls]	[397 cls]	[397 cls, top-10]	1-view→ 2-view
random	1.0000	1.0000	4.00	25.00	0.25	2.52	4.00 → 4.00
CLSNET	0.9239	0.9145	25.11±0.72	41.81±0.38	0.70±0.12	6.10±0.67	-
TEMPORAL [21]	0.7587	0.8119	35.47±0.51	45.12±1.21	1.21±0.14	8.24±0.25	29.60→ 31.90
DRLIM [9]	0.6404	0.7263	36.60±0.41	47.04±0.50	1.02±0.12	6.78±0.32	14.89→ 17.95
EQUIV	**0.6082**	**0.6982**	**38.48±0.89**	**50.64±0.88**	**1.31±0.07**	**8.59±0.16**	**38.52→43.86**
EQUIV+DRLIM	**0.5814**	**0.6492**	**40.78±0.60**	**50.84±0.43**	**1.58±0.17**	**9.57±0.32**	**38.46→43.18**

Table 1. (Left) Average equivariance error (Eq (9)) on NORB for ego-motions like those in the training set (atomic) and novel ego-motions (composite). (Center) Recognition result for 3 datasets (mean ± standard error) of accuracy % over 5 repetitions. (Right) Next-best view selection accuracy %. Our method EQUIV (and augmented with slowness in EQUIV+DRLIM) clearly outperforms all baselines.

Quantitative Results

Tasks→	Equivariance error		Recognition accuracy %				Next-best view
Datasets→	NORB		NORB-NORB	KITTI-KITTI	KITTI-SUN	KITTI-SUN	NORB
Methods↓	atomic	composite	[25 cls]	[4 cls]	[397 cls]	[397 cls, top-10]	1-view→ 2-view
random	1.0000	1.0000	4.00	25.00	0.25	2.52	4.00 → 4.00
CLSNET	0.9239	0.9145	25.11±0.72	41.81±0.38	0.70±0.12	6.10±0.67	-
TEMPORAL [21]	0.7587	0.8119	35.47±0.51	45.12±1.21	1.21±0.14	8.24±0.25	29.60→ 31.90
DRLIM [9]	0.6404	0.7263	36.60±0.41	47.04±0.50	1.02±0.12	6.78±0.32	14.89→ 17.95
EQUIV	**0.6082**	**0.6982**	**38.48±0.89**	**50.64±0.88**	**1.31±0.07**	**8.59±0.16**	**38.52→43.86**
EQUIV+DRLIM	**0.5814**	**0.6492**	**40.78±0.60**	**50.84±0.43**	**1.58±0.17**	**9.57±0.32**	**38.46→43.18**

Table 1. (Left) Average equivariance error (Eq (9)) on NORB for ego-motions like those in the training set (atomic) and novel ego-motions (composite). (Center) Recognition result for 3 datasets (mean ± standard error) of accuracy % over 5 repetitions. (Right) Next-best view selection accuracy %. Our method EQUIV (and augmented with slowness in EQUIV+DRLIM) clearly outperforms all baselines.

Qualitative Results

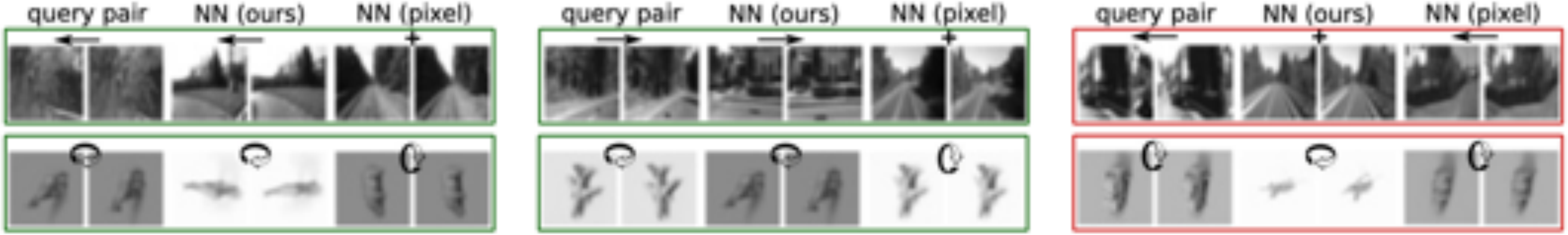


Figure 4. Nearest neighbor image pairs (cols 3 and 4 in each block) in pairwise equivariant feature difference space for various query image pairs (cols 1 and 2 per block). For comparison, cols 5 and 6 show pixel-wise difference-based neighbor pairs. The direction of ego-motion in query and neighbor pairs (inferred from ego-pose vector differences) is indicated above each block. See text.

Contributions

- “Embodied” approach to feature learning that generates features equivariant to ego-motion
- Promising results on multiple datasets and on multiple tasks
  - beneficial for many downstream tasks & other future applications

Questions?