Deep Learning of Invariant Features via Fixations in Video

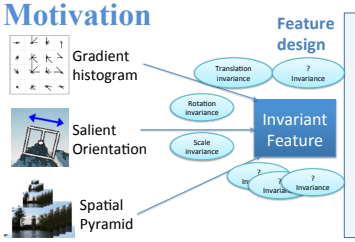
Will Y. Zou, Shenghuo Zhu, Andrew Y. Ng and Kai Yu

NEC Laboratories America
Relentless passion for innovation

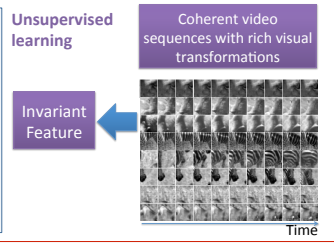
1 Abstract

We apply salient feature detection and tracking in videos to simulate fixations and smooth pursuit in human vision. With tracked sequences as input, a hierarchical network of modules learns invariant features using a temporal slowness constraint. The network encodes invariance which are increasingly complex with hierarchy. Although learned from videos, our features are spatial instead of spatial-temporal, and well suited for extracting features from still images. We applied our features to four datasets (COIL-100, Caltech 101, STL-10, PubFig), and observe a consistent improvement of 4% to 5% in classification accuracy. With this approach, we achieve state-of-the-art recognition accuracy 61% on STL-10 dataset.

2 Motivation



- Hand-designed features encode many type of invariances, and each invariance adds engineering complexity
- Simulated fixations make it possible to extract non-trivial transformations
- We find a natural set of invariances by learning from natural videos

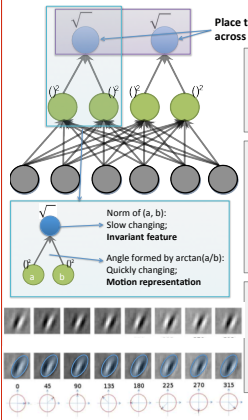


3 Learn hierarchically invariant features, unsupervised

1st layer learning module

$$J = \sum_{t=1}^{N} \|\mathbf{x}^{(t)} - W^T W \mathbf{x}^{(t)}\|_2^2 + \lambda \sum_{t=1}^{N} \|\mathbf{p}^{(t)}\|_1 + \gamma \sum_{t=1}^{N-1} \|\mathbf{p}^{(t)} - \mathbf{p}^{(t+1)}\|_1$$

Reconstruction cost | Sparsity cost | Temporal slowness cost

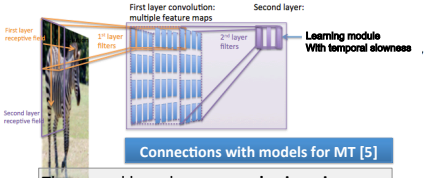


Group size 2 sub-space pooling gives efficient quadrature pair representation

The sub-space structure decomposes representation into 'norm' and 'angle', which gives **a 'what' and 'where' interpretation** [2]

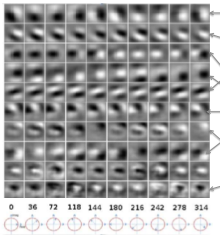
Increment in interpolation angle correspond to **translations** in the local feature

Learning high-level features



Connections with models for MT [5]

The second layer learns **complex invariances.** Visualization of second layer invariances by incrementing subspace pooling angle Second layer features are roughly invariant to these shown transformations



'rotation' like

Large-scale translation

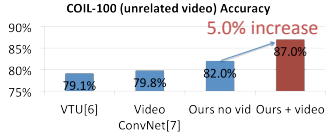
Out-of-plane changes

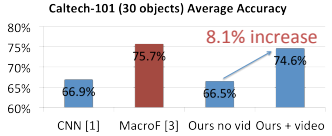
Inward/ outward warping

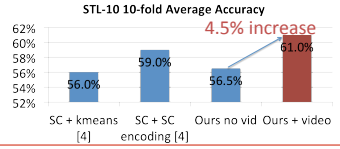
4 Results

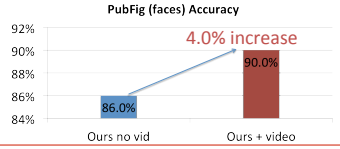
The **features are trained once** on the van Hataren video dataset, and **transferred** to **multiple object recognition datasets**: COIL-100 (87%), Caltech-101 (74.6%), STL-10 (61%), PubFig (90%[our split])

Our approach brings results to competitive or state-of-the-art levels







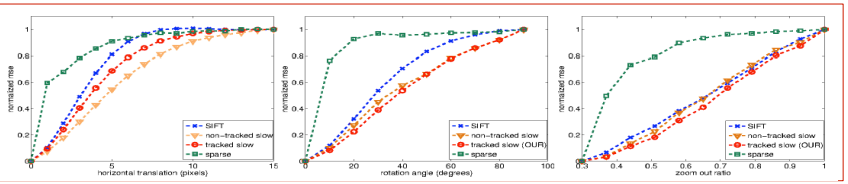


5 Invariance

Translation

Rotation

Zoom



6 References

[1] K. Jarrett, K. Kavukcuoglu, M. Ranzato and Y. LeCun. What is the best Multi-Stage Arch. for Obj. Recog.?. ICCV 2009.
[2] C. Cadieu, B. Olshausen. Learning Transformational Invariants from Natural Movies. NIPS 2010
[3] Y. Boureau, F. Bach, Y. LeCun, and J. Ponce. Learning mid-level features for recognition. CVPR 2010
[4] A. Coates, A. Y. Ng. The Importance of Encoding Versus Training with Sparse Coding and Vector Quantization. ICML 2011
[5] N. Rust, V. Mante, E. P. Simoncelli, and J. A. Movshon. How mt cells analyze the motion of visual patterns. Nature Neuroscience, 2006
[6] H. Wersing and E. Kroner. Learning optimized features for hierarchical models of invariant object recognition. Neural Comp. 2003
[7] H. Mobahi, R. Collobert and J. Weston. Deep learning from temporal coherence in video. ICML 2009

http://ai.stanford.edu/~wzou/