Importance of Models in Vision:
Hierarchical Compositional Representations

Aleš Leonardis

University of Birmingham
School of Computer Science
Centre for Computational Neuroscience
& Cognitive Robotics

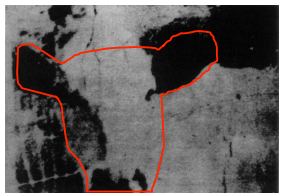
UNIVERSITY OF BIRMINGHAM

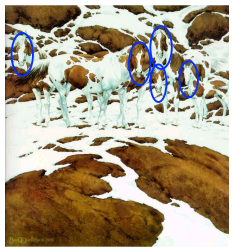
CNCR

Outline

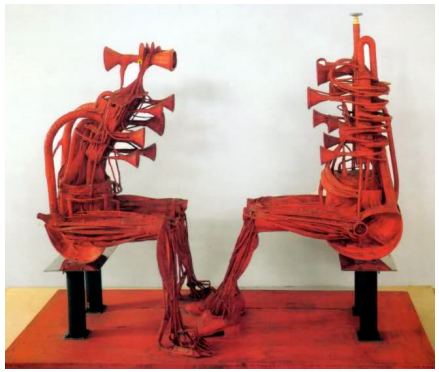
- Motivation (Why is vision a hard problem?)
- Vision as an ill-posed problem
- Regularization
- Models: a historical perspective
- Requirements for models
- Hierarchical compositional representations
- Extensions, generalizations
- Conclusions

Why is vision a hard problem?

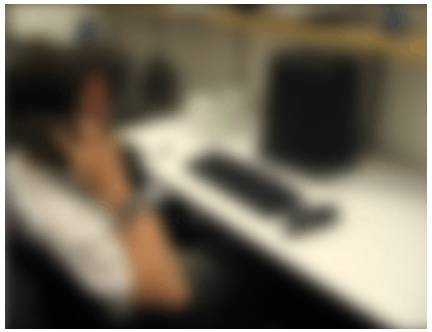




Why is vision a hard problem?



Why is vision a hard problem?



Slide credit: A. Torralba

Why is vision a hard problem?

Slide credit: A. Torralba

Why is vision a hard problem?

A Physician Riding a Donkey, by Niko Pirosmanashvili

You Who Can't Do Anything, by Francisco Goya

Image understanding

"A harbor with many dozens of boats; water is calm and glassy; masts are all vertical; mountains in background, blue sky with a touch of clouds…"

Slide credits: M Turk

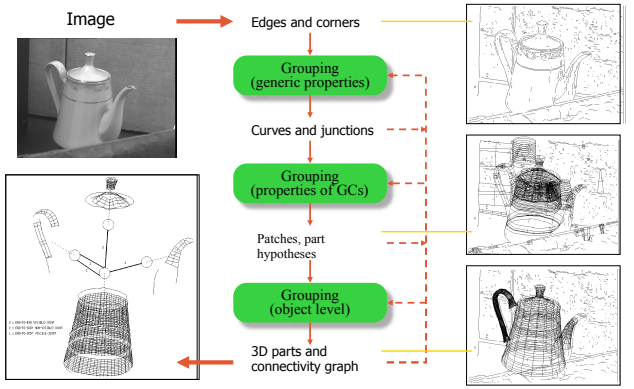
Human perception

- Inference of contours when there is no contrast in the image
- Inference of objects in clutter (occlusions, missing information)
- Inference of 3D shape from 2D contours, from textured patterns
- Inference of motion from static images
- Inference of a true surface color under different lighting conditions

- *Perception is a kind of controlled hallucination [Max Clowes, Jan Koenderink]*

- Vision is an ill-posed problem which requires regularization.

From image to 3-D description



G. Medioni, *Generic shape learning and recognition*, Workshop on Generic Object Recognition and Categorization (CVPR 2004)

Detecting a glass

Is a general computer vision possible?

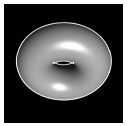
- Reliable results can be obtained
  - when data is well-behaved (follows the assumptions)
  - when we have strong models (can compensate for missing data, noise, deviations from assumptions)

- Is a general computer vision possible?
  - data is not well behaved (no purely bottom-up, signal-based approach will ever work)
  - Ill-posed problems, general regularization approaches fail
  - strong models are needed (prior, knowledge): memory-based vision (prohibitive complexity?)

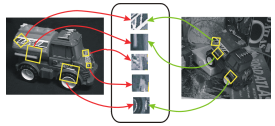
Models

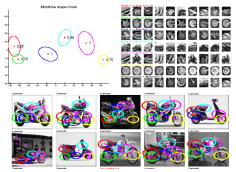
- Blocks world
- Polyhedral models
- Geons/superquadrics
- Regularized data using different functionals
- Statistical models



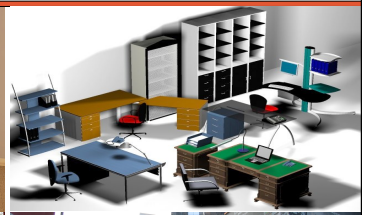


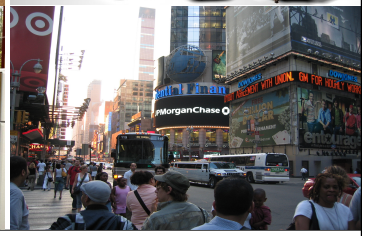




Large number of visual object classes

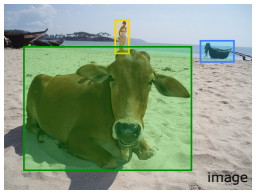


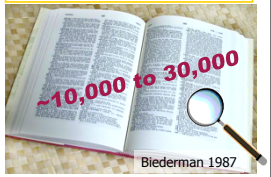
Large number of visual object classes

"Now! *That* should clear up a few things around here!"

Intra-class variability, articulations,…

- A large number of object classes
- Significant intra/inter-class variation
- Multiple articulations
- Multiple 3D poses
- Varying illuminations
- Objects can appear at any position in an image, any scale, orientation…





Tasks

- Recognition of exemplars

- Categorization
  - Subordinate-
  - Basic-
  - Super-ordinate-level categories

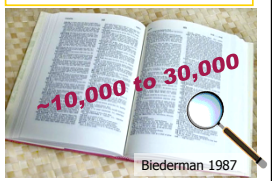
Tasks

- Grasping
- Manipulation
- Talking and reasoning about objects

Central issues

Central issues:

- **Representation**
- **Inference**
- **Learning**



Requirements for good representations

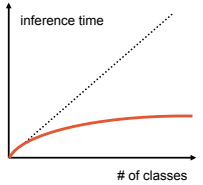
- Representations, inference and learning: the key issues
- Requirements:

A Representation should:

- Be generative (robustness): also, support a variety of tasks
- Enable fast and robust (object) detection/segmentation/ parsing
- Scale with the number of classes (modest increase in memory)
- Accommodate exponential variability (of objects)
- Enable efficient learning

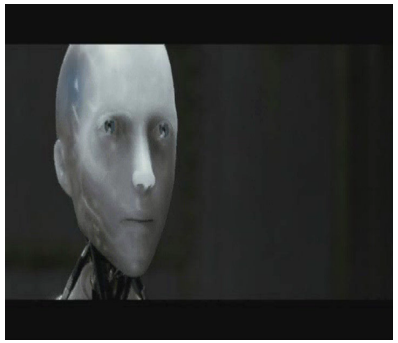
Requirements

- Inference
  - Sub-linear in the number of classes
  - Coping with noisy or missing information (make predictions)



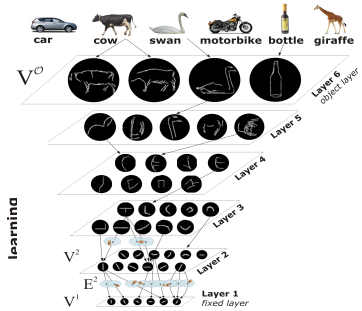
- Learning should:
  - Require minimal human effort
  - Be done incrementally, on-line (no need for re-training the complete representation)
  - Share-ability (in terms of representation and processing)
  - Transfer of knowledge (learning time getting shorter)
  - Scaffolding (gradual increase of knowledge)

Our Approach

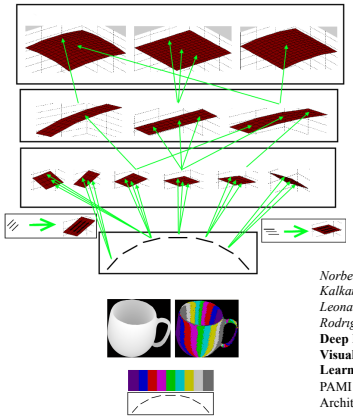


Compositional hierarchies

- A 2D shape hierarchy
  - Representations
  - Inference
  - Learning



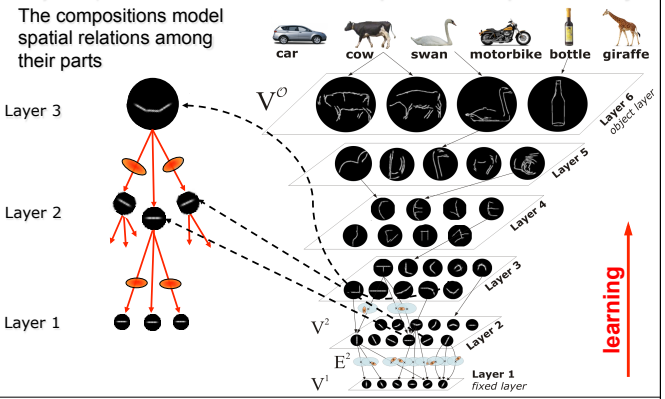
3D compositional shape hierarchy



Norbert Kruger, Peter Janssen, Sinan Kalkan, Markus Lappe, Ales Leonardis, Justus Piater, Antonio J. Rodriguez-Sanchez, Laurenz Wiskott, **Deep Hierarchies in the Primate Visual Cortex: What Can We Learn For Computer Vision?** IEEE PAMI 2013, SI: Learning Deep Architectures

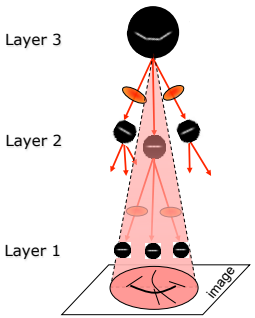
Our approach - representation

- Object representation: A **hierarchical compositional shape vocabulary**
- The compositions model spatial relations among their parts

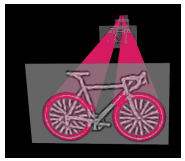


Our approach - representation

- Object representation: A **hierarchical compositional shape vocabulary**
- The compositions model spatial relations among their parts

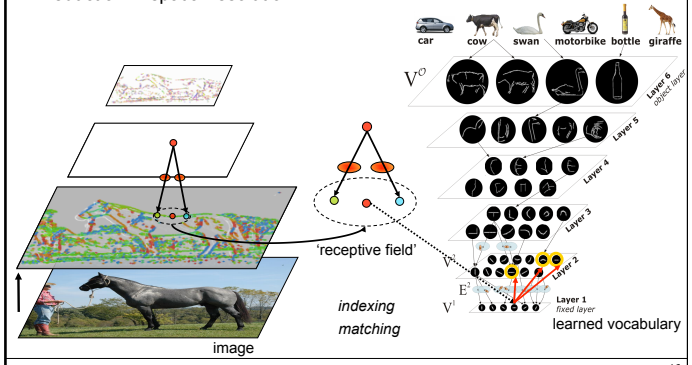


- Invariance to local deformations
- Exponential flexibility
- Robustness to clutter
- Fast inference



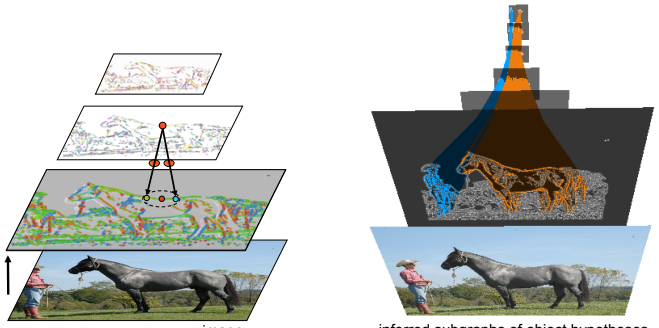
Inference

- Inference proceeds bottom-up
- Reduction in spatial resolution
- Indexing and matching



Inference

- 700x500 images
- Detecting multiple categories

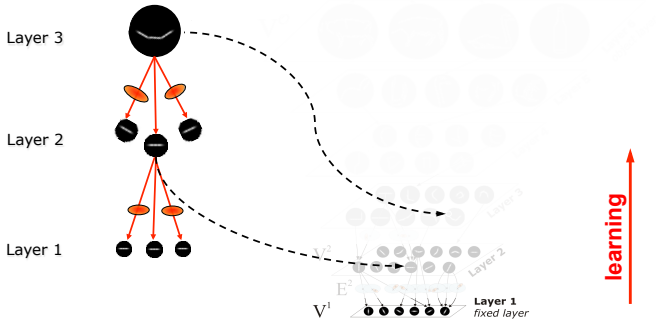


image

inferred subgraphs of object hypotheses

Learning

- **Bottom-up**

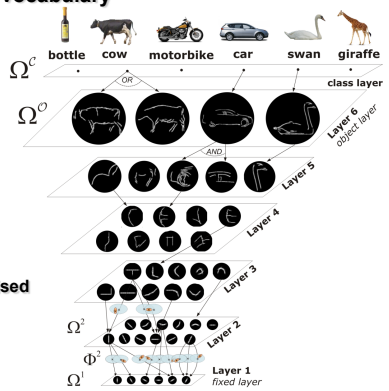


Learning

- **Learning the hierarchical vocabulary**
  - Learn the **number** of compositions at each layer
  - Learn the **structure** of each composition (the number of parts and the parameters of the distributions)

Learning of **structure** is **unsupervised**

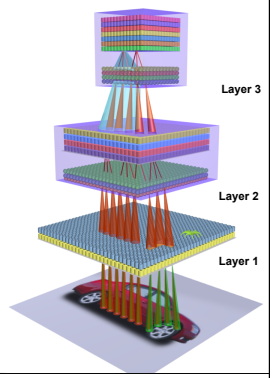
Learning of **classes** is **supervised**



Compositional shape hierarchy

- A Computational Model for Learning a Multi-Level Compositional Representation of Visual Structure

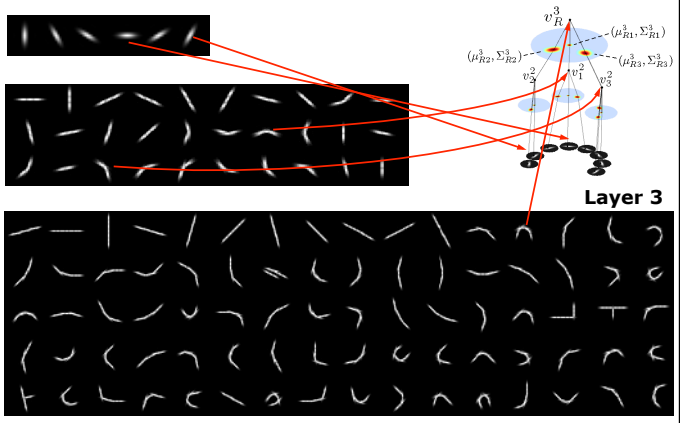
  - Computational plausibility
    - Hierarchical representation
    - Compositionality (*parts composed of parts*)
    - Indexing & matching recognition scheme
  - Statistics driven learning (unsupervised learning)
  - Fast, incremental (continuous) learning



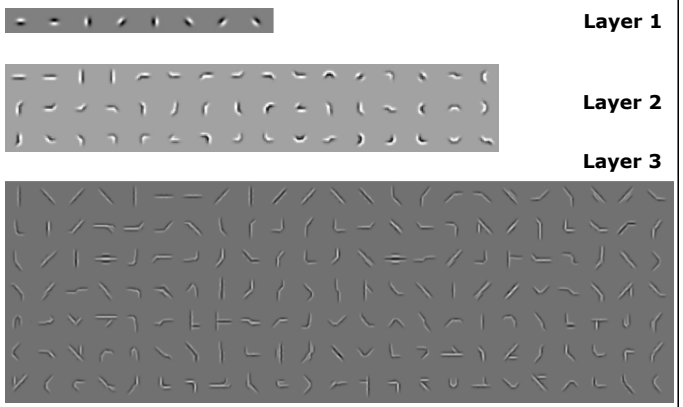
Experimental results

- Learning a vocabulary from:
  - a set of natural images
  - a set of “Gaussian noise” images
  - a set of “letters” images
- starting from
  - a set of oriented edges
  - a set of polarity edges
  - DOG / on-off cells
- Multi-class object detection
- Share-ability, transfer of knowledge, incremental learning
- Scalability -> Taxonomy of object categories

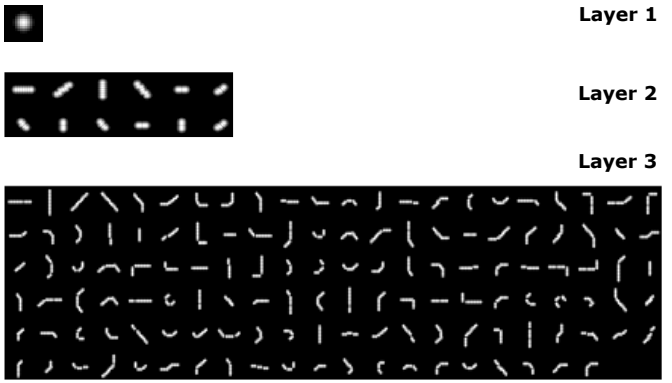
Natural images, edge filters



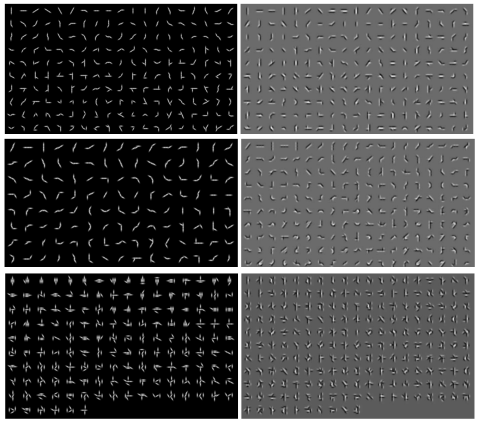
Natural images, polarity filters



Natural images, DoG filters



Natural images, edge filters



Natural images, edge filters

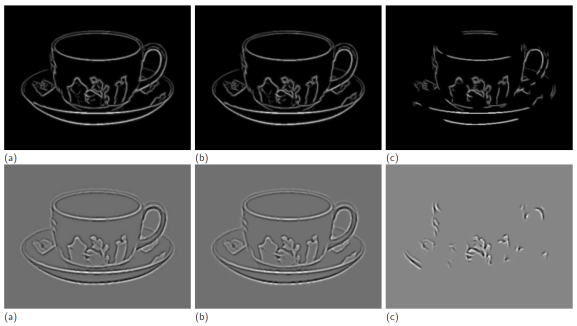
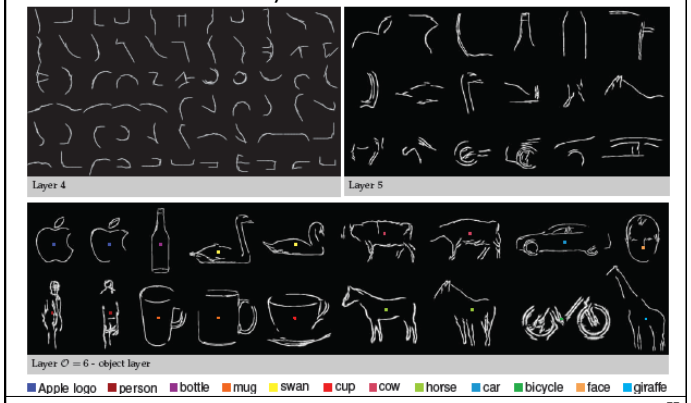


Fig. 9. The image points that Layer 2 vocabularies "see": (a) natural objects, (b) faces, (c) Gaussian noise. Top row: edge filters, bottom row: polar filters.

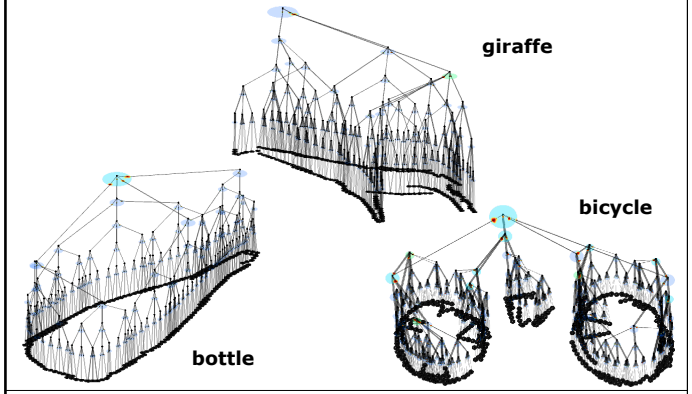
Multi-class learning and detection

- Learned vocabulary

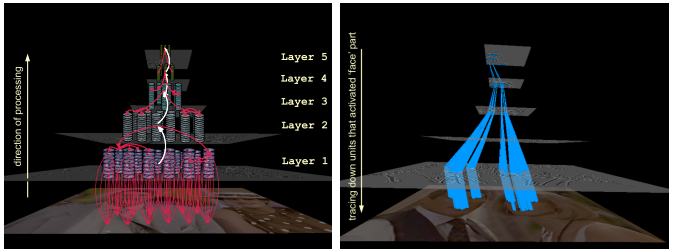


Multi-class learning and detection

- Examples of learned whole-object shape models



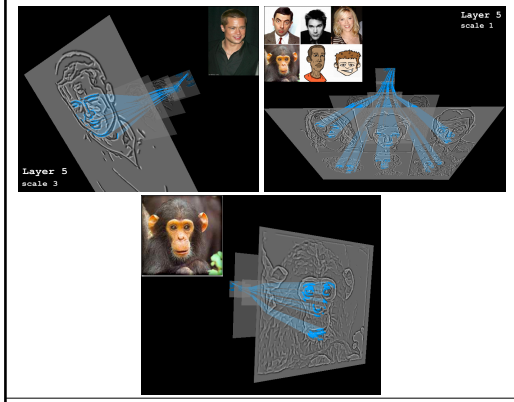
Detection



Inference proceeds bottom-up. Active parts can easily be “traced” down to the image.

Object detection and recognition

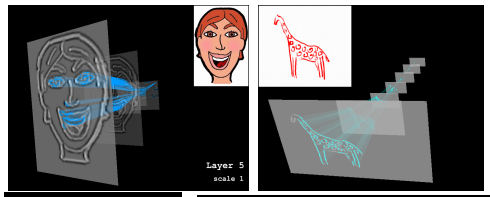
- **Invariance**



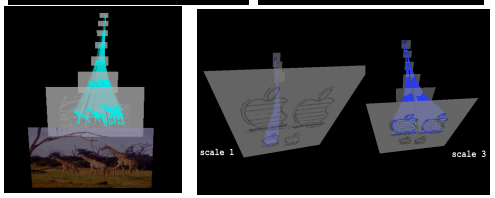
- intra-class variability

Object detection and recognition

- **Invariance**

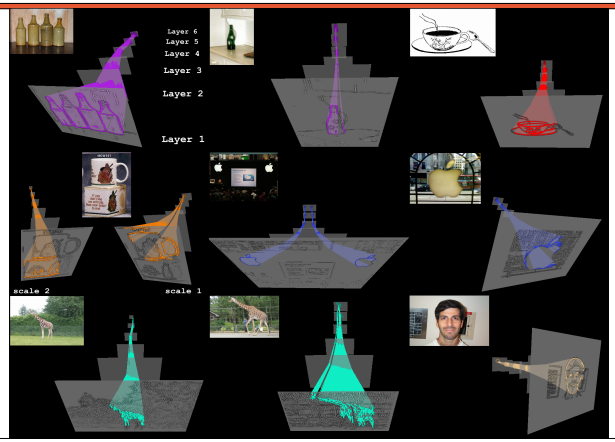


- real / hand drawn

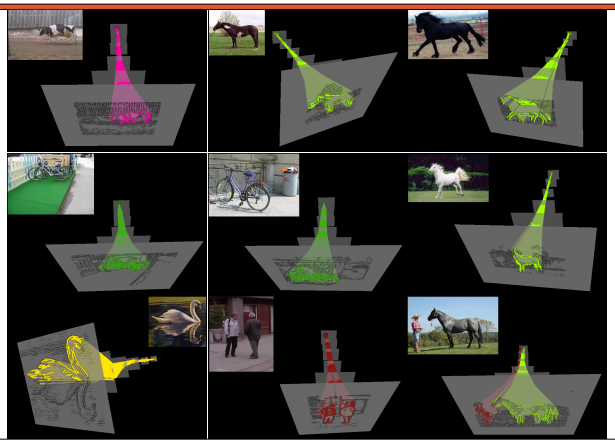


- scale

Detection of multiple object classes



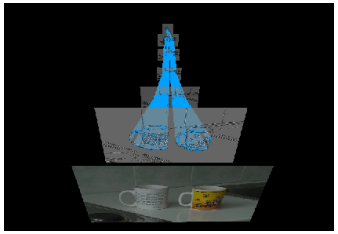
Detection of multiple object classes



Object detection and recognition

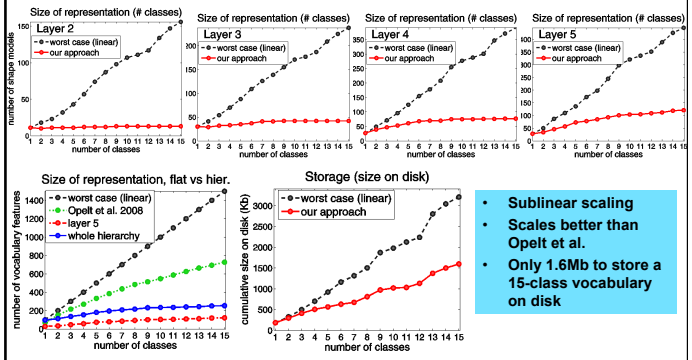
- Detecting unknown categories, triggering learning

Detection of object classes, cups



Size of the vocabulary

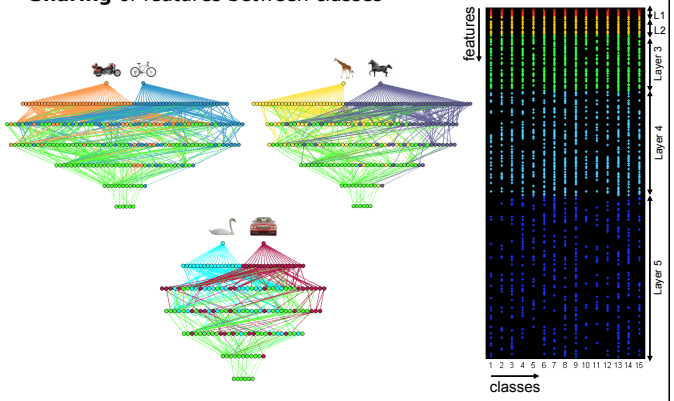
- **Size of the vocabulary** as a function of the number of learned class



- **Sublinear scaling**
- **Scales better than Opelt et al.**
- **Only 1.6Mb to store a 15-class vocabulary on disk**

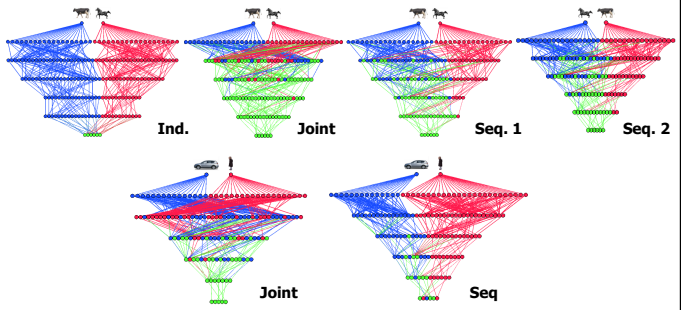
Sharing of features

- **Sharing** of features between classes

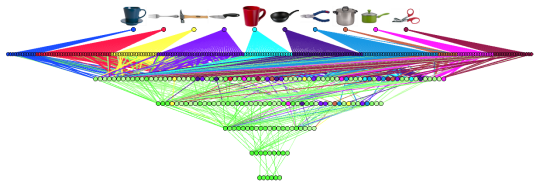


Multi-class learning strategies

- **Feature sharing** among similar and dissimilar classes
  - Joint achieves the best sharing of features. Sequential is comparable.
  - Sharing is also present for visually dissimilar objects (lower layers)

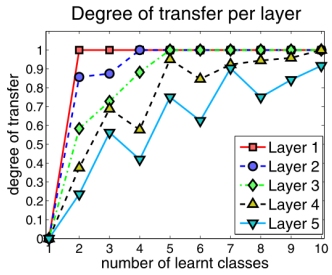


Sharing of features



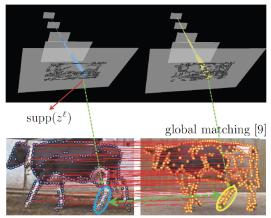
Transfer of features

- **Transfer** of features in incremental learning



Shape consistency and deformations

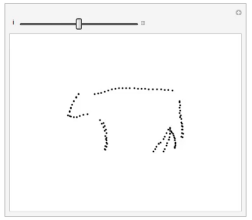
- Example: putting two compositions (blue and yellow) representing a leg into correspondence by global matching of two cows.



Two compositions are matched, if the global matching maps supports of the two compositions one to another (significant portion of them).

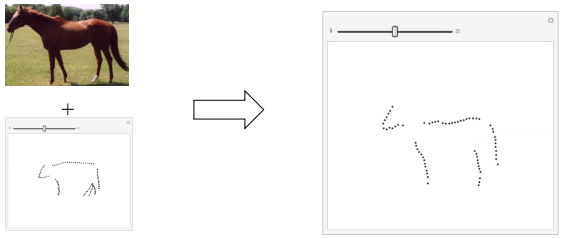
Learning shape consistency and deformations

- Examples: Deformations/articulations of a cow model.



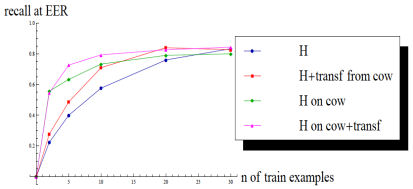
Transfer of deformations

- Transfer of deformations to novel classes:
  - Example: transfer of variation of cow parts to one horse training image



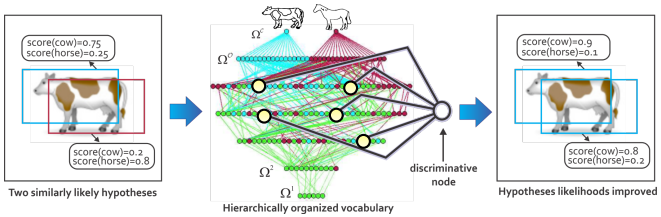
Transfer of deformations

- Transfer of deformations to novel classes:
  - Results: Recall at EER for horses at different number of training examples by borrowing from cows



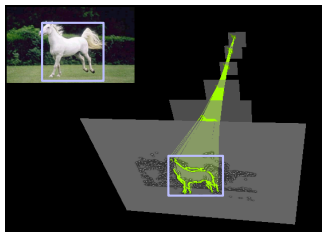
Adding discriminative power

- Goal: Identify subset of parts and combine them into a discriminative node to improve discrimination.

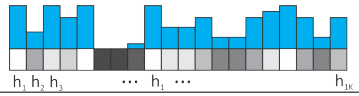


Adding discriminative power

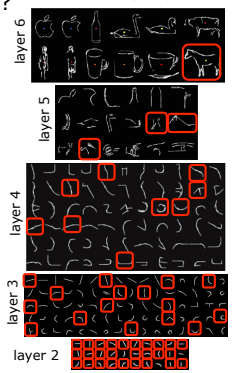
- Which parts activate at detection?



Cumulative histogram of responses over parts:

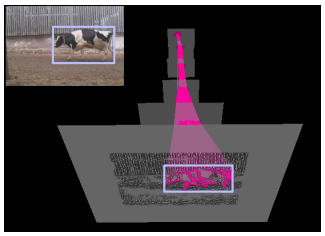


The library of parts

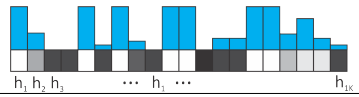


Adding discriminative power

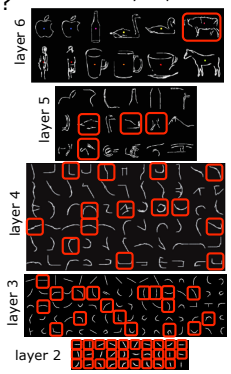
- Which parts activate at detection?



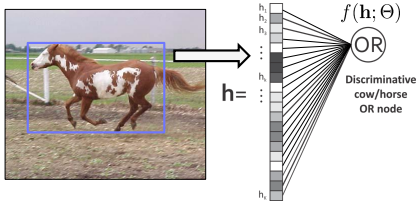
Cumulative histogram of responses over parts:



The library of parts



Adding discriminative power

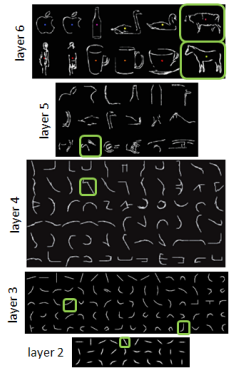


$$f(\mathbf{h};\Theta) = \hat{\mathbf{h}}^T\Theta$$

Linear combination of parts responses

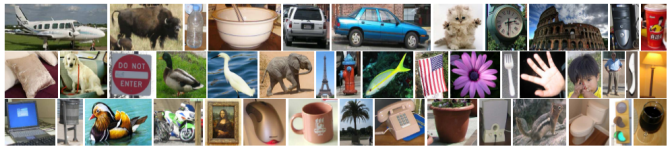
Sparse solutions of Θ identify discriminative parts in our library!

The library of parts



Large scale experiments

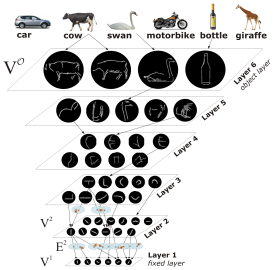
- Experiment on LabelMe dataset

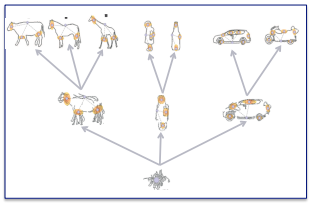


- 175 object classes: shape as main clue:
  - car in different 3D views, motorbike, mug, bottle, orange, street-lamp, window, person, bird, laptop, mouse, …
- 30 examples per class

Feature hierarchies

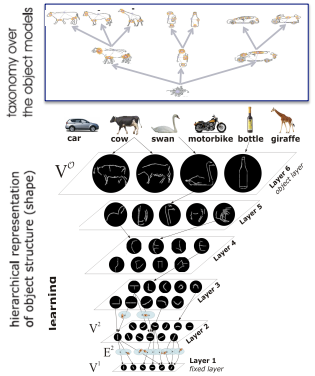
- Sharing features among object classes; logarithmic increase in memory storage
- Top, object, layer: **linear**; problem with many object classes





Final Representation

- Final representation: TCT (taxonomy of object classes) + LHOP (hierarchy of shape features)
- Logarithmic in the size of the shape of each object class
- Efficient (logarithmic?) in the number of object classes
- Adding discriminative nodes

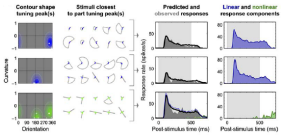


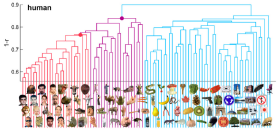
Summary and discussion

- Computational principles towards building complex representations
- Scaling in terms of memory, speed-up of inference, efficient learning
- General insights
  - Modeling/memorizing large-scale spatial-temporal patterns
    - Other modalities
    - Other senses
    - Sensing as a "controlled hallucination"

Work in progress

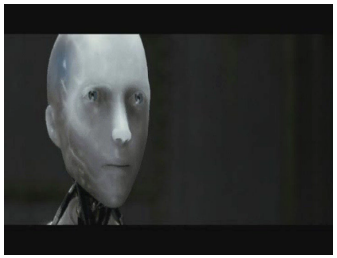
- relate parts to 3D concepts / affordances / grasping modes,
- relate parts to actions,
- relate (semantic) parts to words,
- attention, context
- hierarchical compositionality for sound, music, speech, touch, motor system
- relations to biology





Norbert Kruger, Peter Janssen, Sinan Kalkan, Markus Lappe, Ales Leonardis, Justus Piater, Antonio J. Rodrıguez-Sanchez, Laurenz Wiskott, **Deep Hierarchies in the Primate Visual Cortex: What Can We Learn For Computer Vision?** IEEE PAMI 2013, SI: Learning Deep Architectures

Thank you



Thanks to Marko Boben, Matej Kristan, Sanja Fidler, Domen Tabernik, Vladislav Kramarev

The work was supported in part by EU projects: Cosy, Poeticon, Mobvis, CogX, PaCMan, USA DARPA project: Neovision2; National projects: ARRS.