Compositional hierarchies for scalable robot vision

Aleš Leonardis

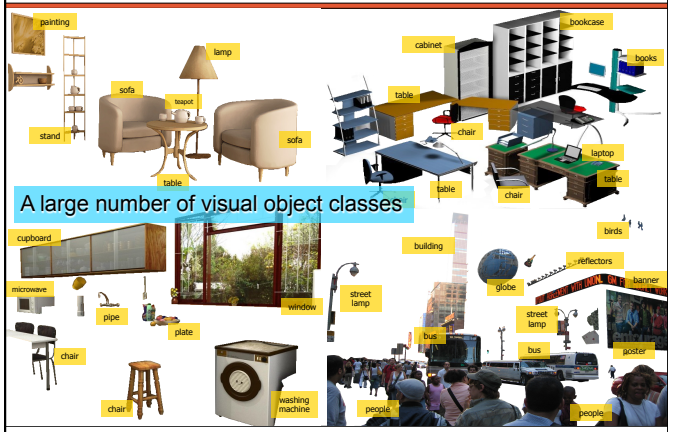
University of Birmingham
School of Computer Science
Centre for Computational Neuroscience
& Cognitive Robotics

Outline

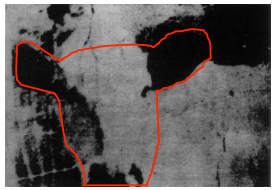
- Motivation – different faces of scalability
  - large number of object categories / means of regularization
  - in terms of processing (learning, inference)
  - in terms of dealing with multiple tasks
- Requirements for a representation, inference, learning
- Hierarchical compositional representations
  - 2D shape
  - incremental learning, transfer of knowledge
  - generative-discriminative
  - multiple tasks
  - 3D shape
- Conclusions

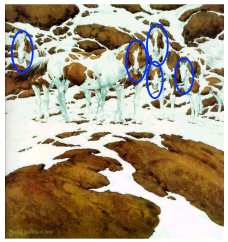
Large number of visual object classes

Large number of visual object classes



Strong models for regularization





Strong models for regularization

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

Tasks

- Recognition of exemplars

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

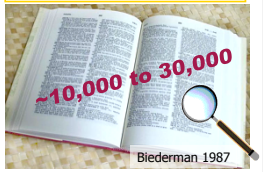
Tasks

- Grasping
- Manipulation
- Talking and reasoning about objects

Central issues

Central issues:

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



Requirements for good representations

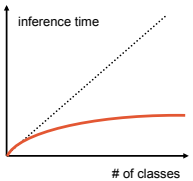
- Representations, inference and learning: the key issues
- Requirements:

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

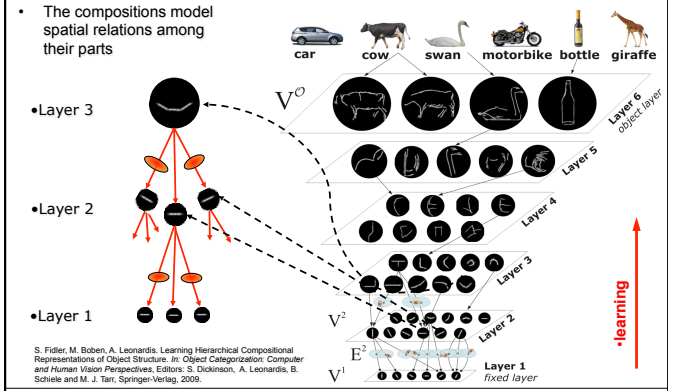
- Inference
  - Sub-linear in the number of classes
  - Coping with noisy or missing information (make predictions)



- 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)

Compositional hierarchy

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



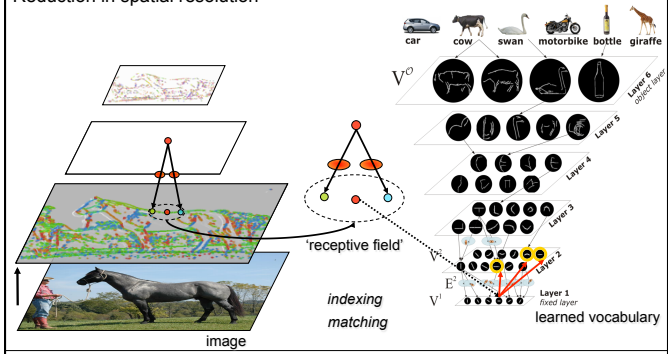
S. Fidler, M. Boben, A. Leonardis. Learning Hierarchical Compositional Representations of Object Structure. *In: Object Categorization: Computer and Human Vision Perspectives*, Editors: S. Dickinson, A. Leonardis, B. Schiele and M. J. Tarr, Springer-Verlag, 2009.

Inference

Inference proceeds bottom-up

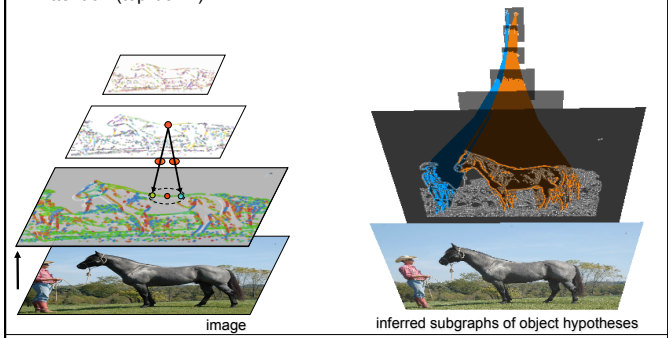
Reduction in spatial resolution

Indexing and matching



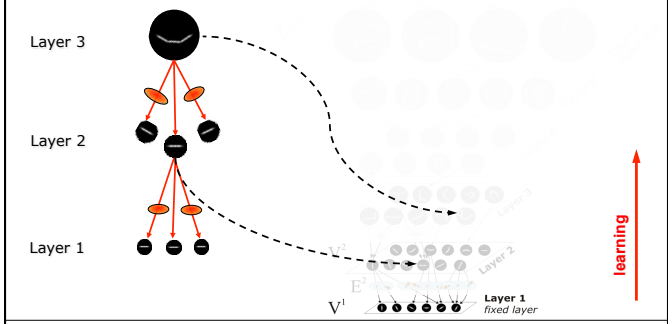
Inference

- Detecting multiple categories
- Prediction (bottom-up)
- Attention (top-down)



Learning

- **Bottom-up**

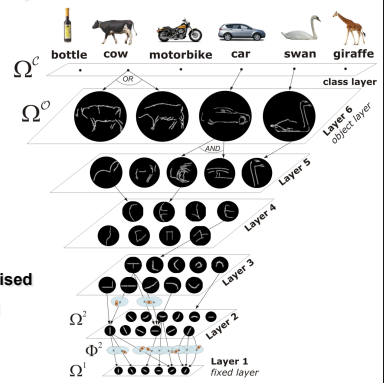


Learning

- **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)

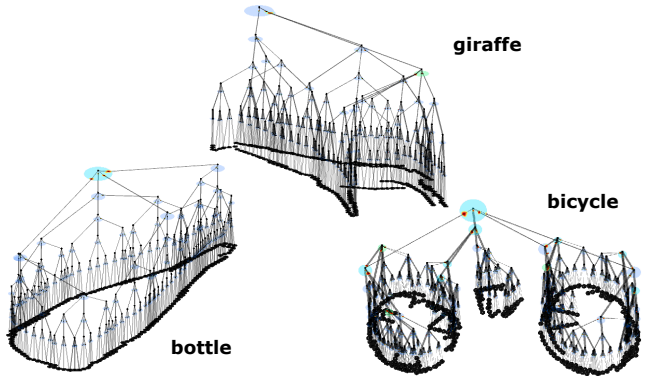
Learning of **structure** is **unsupervised**

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

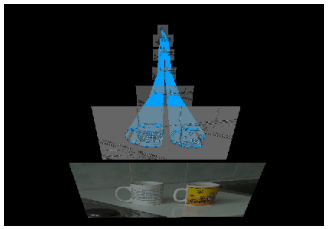


Multi-class learning and detection

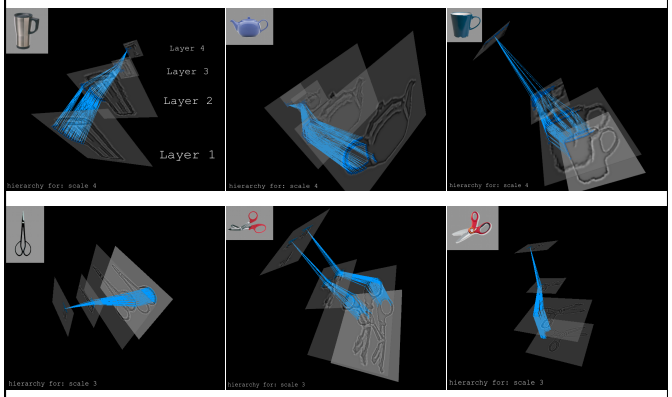
- Examples of learned whole-object shape models



Detection of object classes, cups

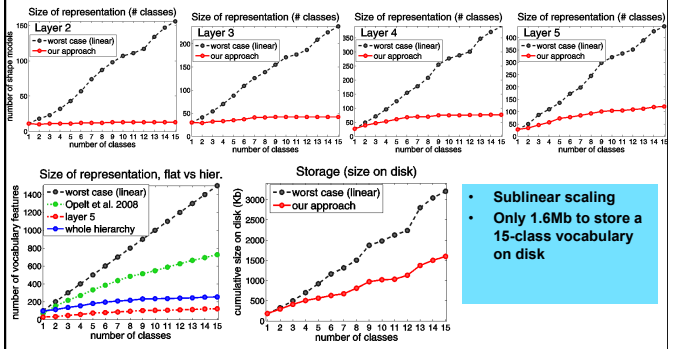


Detection at multiple layers



Size of the vocabulary

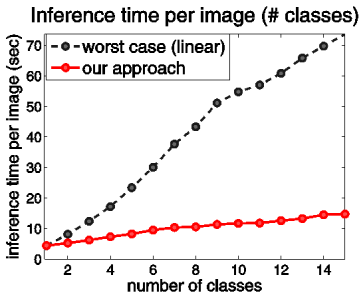
- **Size of the vocabulary** as a function of the number of learned class



- **Sublinear scaling**
- **Only 1.6Mb to store a 15-class vocabulary on disk**

Inference time

- **Inference time** (average per image) as a function of the number of learned class

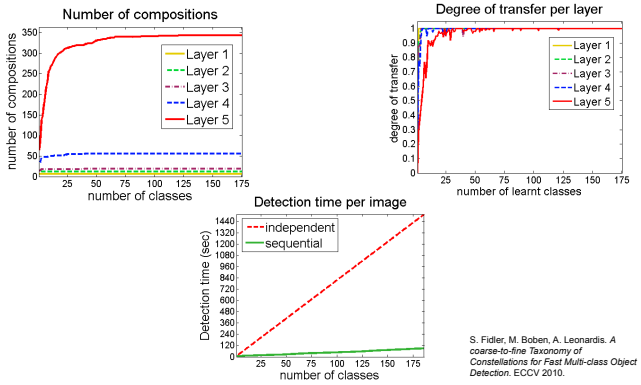


Hardware information:
- Intel Xeon-4 CPU 2:66 Ghz computer (one core used)
- implemented in C++

- Only 16 seconds per image (approx. 500x700) for 15-class object detection

Large scale experiments

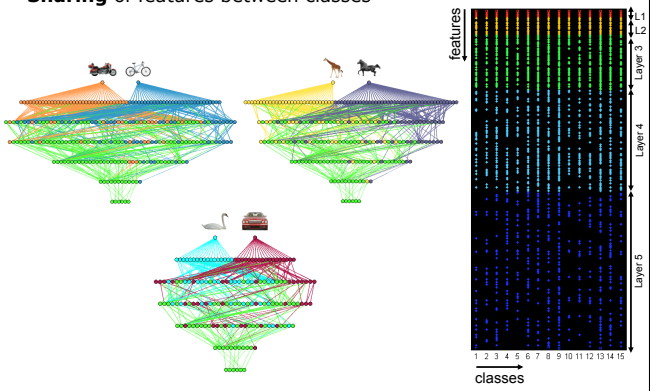
- **Large-scale experiments**



S. Fidler, M. Boben, A. Leonardis. *A coarse-to-fine Taxonomy of Constellations for Fast Multi-class Object Detection*. ECCV 2010.

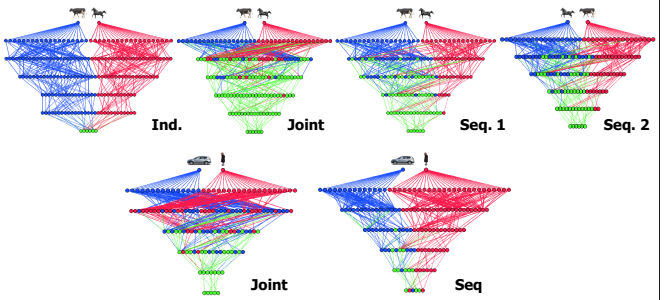
Sharing of features

- **Sharing** of features between classes



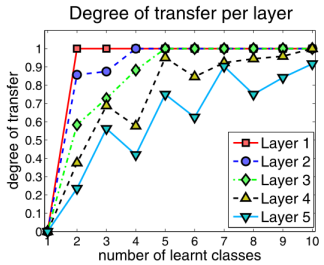
Multi-class learning strategies

- **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)



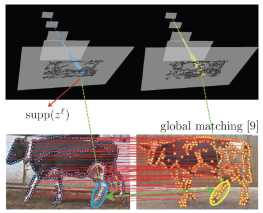
Transfer of features

- **Transfer** of features in incremental learning



Shape consistency and deformations

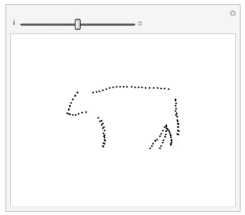
- Example: putting two compositions (blue and yellow) representing a leg into correspondence by global matching of two cows.



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

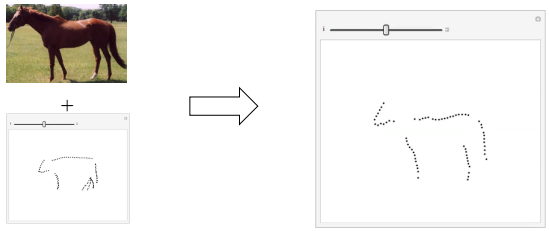
Learning shape consistency and deformations

- Examples: Deformations/articulations of a cow model.



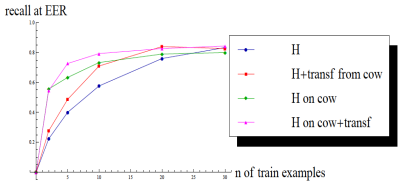
Transfer of deformations

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



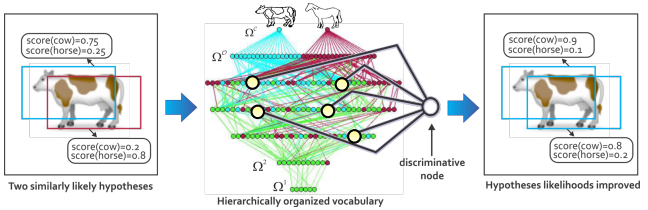
Transfer of deformations

- 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

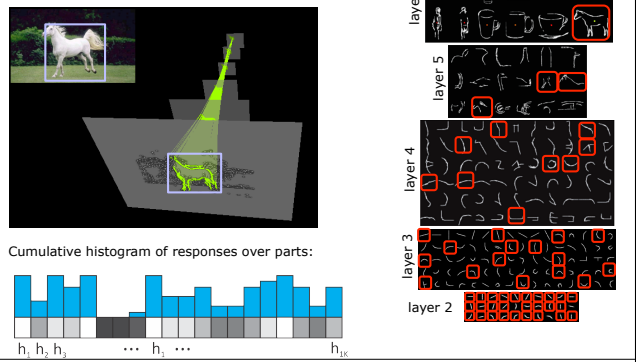
- Goal: Identify subset of parts and combine them into a discriminative node to improve discrimination.



M. Kristan, M. Boben, D. Tabernik, and A. Leonardis, Adding discriminative power to hierarchical compositional models for object class detection, 18th Scandinavian Conference on Image Analysis, SCIA, 2013

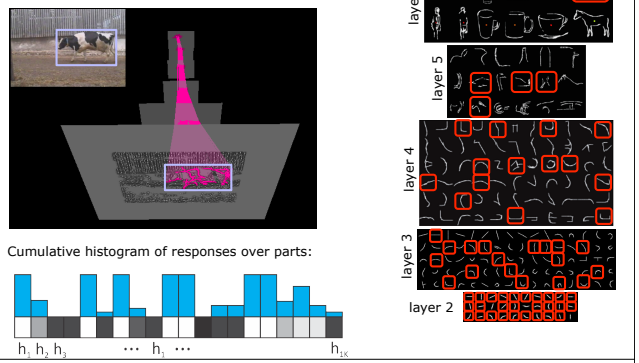
Adding discriminative power

- Which parts activate at detection?

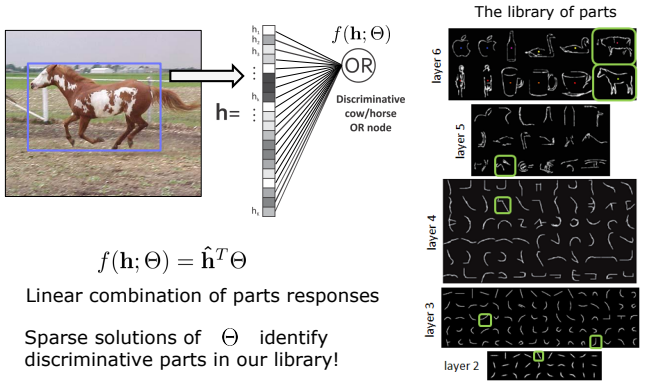


Adding discriminative power

- Which parts activate at detection?



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!

Results

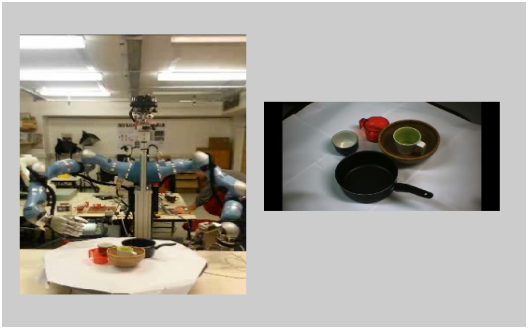
Table 2. Hypothesis voting and ranking stage detection rates using the Pascal 50% overlap criterion on ETHZ [3] at FPPI=1.0. The N_{disc} denotes the number of discriminative nodes selected by dlHoP along with standard deviation in brackets.

	lHoP [12]	dlHoP [N_{disc}] **our work**	PSM [21]	Hough [3]	w_{ac} [20]	M^2HT [19]	PMK [20]	PMK [21]
Apple	92.5	92.5 [5.2 (1.3)]	90.4	43.0	80.0	85.0	80.0	90.4
Bottle	79.6	85.4 [7.4 (1.7)]	84.4	64.4	92.4	67.0	89.3	96.4
Giraffe	75.1	82.3 [13 (4.6)]	50.0	52.2	36.2	55.0	80.9	78.8
Mug	85.9	86.5 [13.2 (6.9)]	32.3	45.1	47.5	55.0	74.2	61.4
Swan	58.6	70.5 [6 (2.6)]	90.1	62.0	58.8	42.5	68.6	88.6
Average	78.3	**83.4 [9.0 (5.1)]**	69.4	53.3	63.0	60.9	78.6	83.2

M. Kristan, M. Boben, D. Tabernik, and A. Leonardis, Adding discriminative power to hierarchical compositional models for object class detection,18th Scandinavian Conference on Image Analysis, SCIA, 2013

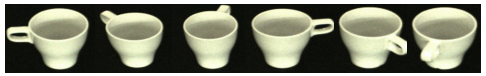
Multiple tasks

A camera-robot setup for data acquisition



Exemplars, categories, multiple views

- Generative-Discriminative learning across multiple views of
  - a single object

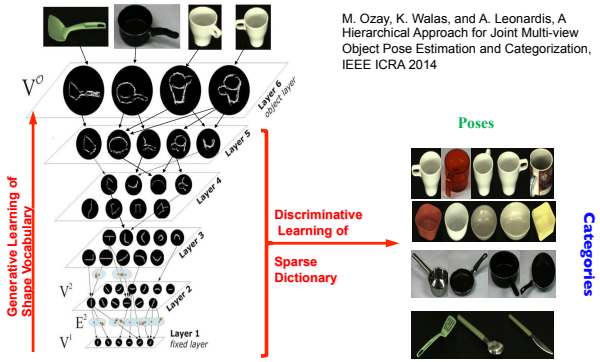


  - multiples instances of objects belonging to a single category,

  - multiples instances of objects belonging to multiple categories

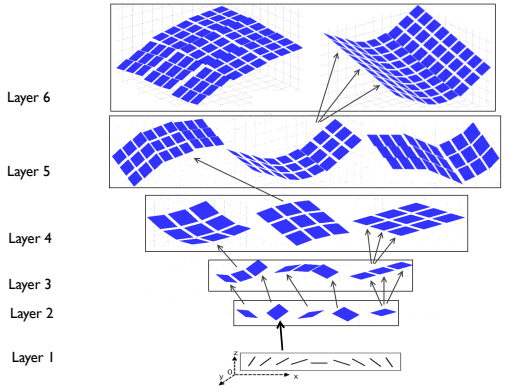
Exemplars, categories, multiple views

- Generative-Discriminative learning by integrating information extracted from multiple layers:

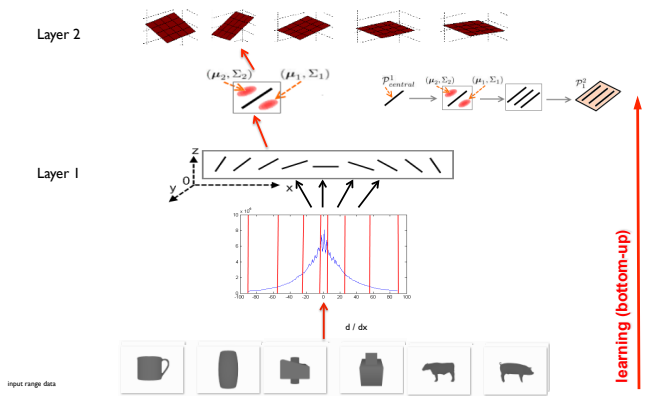


M. Ozay, K. Walas, and A. Leonardis, A Hierarchical Approach for Joint Multi-view Object Pose Estimation and Categorization, IEEE ICRA 2014

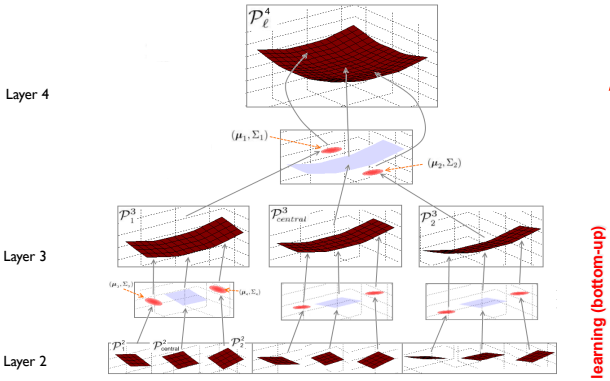
3D Compositional Hierarchy (Representation)



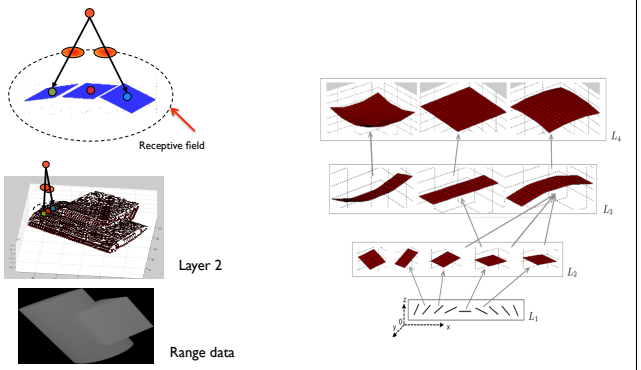
Learning 3D compositional hierarchy (Layers 1, 2)



Learning 3D compositional hierarchy (Layers 3, 4)



3D compositional hierarchy (Inference)

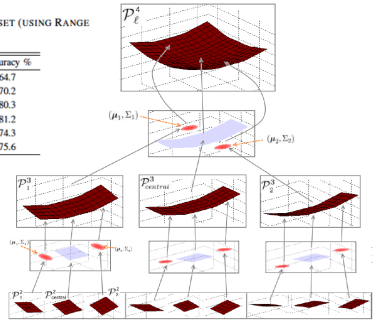


Object Categorization from Range Images

V. Kramarev, S. Zurek, J. L. Wyatt, A. Leonardis: Object Categorization from Range Images using a Hierarchical Compositional Representation, IEEE ICPR 2014

TABLE II. RESULTS FOR RGB-D OBJECT DATASET (USING RANGE IMAGES ONLY)

Method	Accuracy %
Spin Images & 3D Bounding Boxes	64.7
Sparse Distance Learning	70.2
RGB-D Kernel Descriptors	80.3
Hierarchical Matching Pursuit	81.2
Our method (up to L_3)	74.3
Our method (up to L_4)	75.6



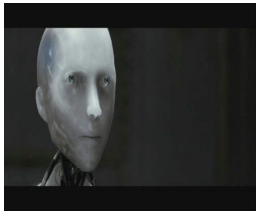
Summary

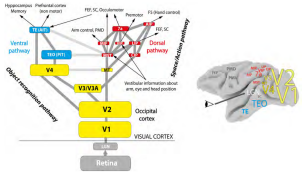
- Computational principles towards building complex representations
- Scaling in terms of memory, speed-up of inference, efficient learning
- General insights
  - Modeling/memorizing large-scale spatial-temporal patterns
    - Other modalities
    - Other senses
    - Sensing as a "controlled hallucination"

Discussion and open questions

- The emphasis should be on firm theoretical models of computational complexity being grounded in statistics of visual object categories
- A new challenge related to scalability
  - in a similar spirit as The Pascal Visual Object Classes challenge but being geared towards efficient large scale learning, low storage and efficient sub-linear inference; and also related to various (cognitive) tasks.

Thank you





N. Krüger, P. Janssen, S. Kalkan, M. Lappe, A. Leonardis, J. H. Piater, A. J. Rodríguez-Sánchez, L. Wiskott: Deep Hierarchies in the Primate Visual Cortex: What Can We Learn for Computer Vision? IEEE Trans. Pattern Anal. Mach. Intell. 35(8): 1847-1871 (2013).

Thanks to Marko Boben, Matej Kristan, Sanja Fidler, Domen Tabernik, Mete Ozay, Rusen Aktas, Vladislav Kramarev

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

Thank you!