Adding discriminative power to hierarchical compositional models for object class detection

University *of Ljubljana*
Faculty *of Computer and Information Science*

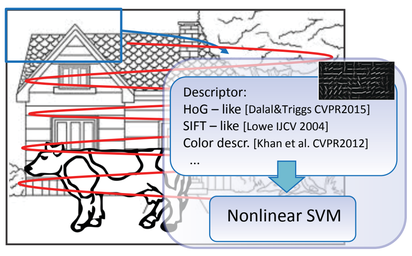
Matej Kristan[(1)], Marko Boben[(1)], Domen Tabernik[(1)], Aleš Leonardis[(2,1)]
[(1)]CIS, University of Ljubljana, [(2)] CN-CR Centre, University of Birmingham
{matej.kristan}{marko.boben}{domen.tabernik}{ales.leonardis}@fri.uni-lj.si

ViCoS visualgnitive systemslab

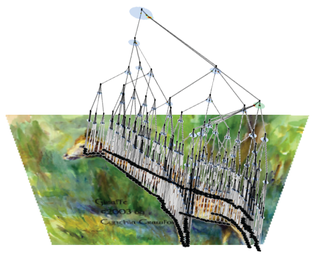
1 Standard vs. hierarchical object detection

- **Standard** object category **detection** = sliding windows or oversegmentations + high-dimensional descriptors + SVM.
- **Problems**: **slow** inference time, **large** storage requirement, **poor** scaling.
- **Hierarchical compositional models** (e.g., [1,2,3,4]) can avoid these drawbacks.
  - Hierarchical: organize the representations by level of granularity.
  - Compositional: each part is a composition of parts from a lower layer.

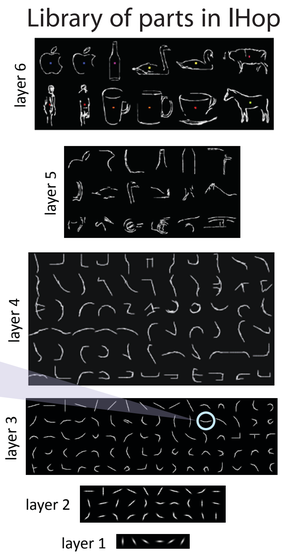
A sliding-window inference



Hierarchical inference

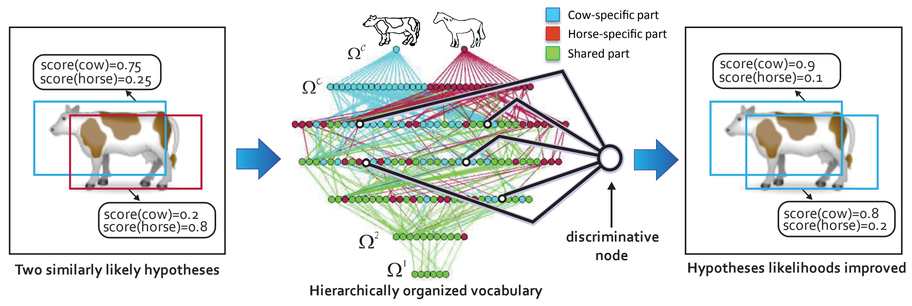


Library of parts in lHop



2 Issues with generative hierarchies

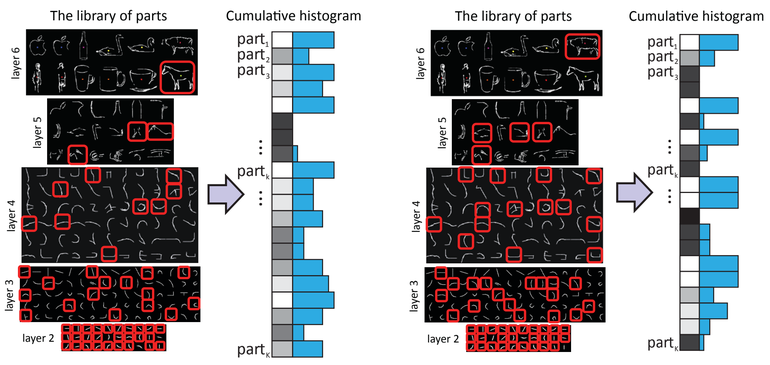
- Layered hierarchy of parts (lHop) [1]
  - **Fast** inference, **good** scaling and **low** storage.
  - Due to significant sharing of parts acrosss categories.
- **Problem:** difficult to discriminate between **visually-similar categories.**
- **Proposed solution**: identify the subset of parts that differentiate pairs of categories and combine them into **discriminative nodes**.



2 Which parts are active at detection?

- Detection of a horse (left) and of a cow (right) by lHop [1]:

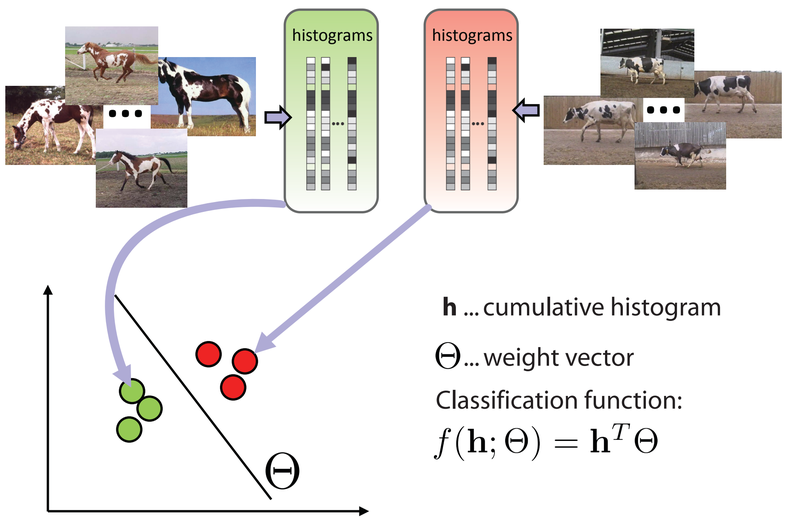
- The frequency/strength of the activated parts can be sumarized by a **cumulative histogram** of part responses:



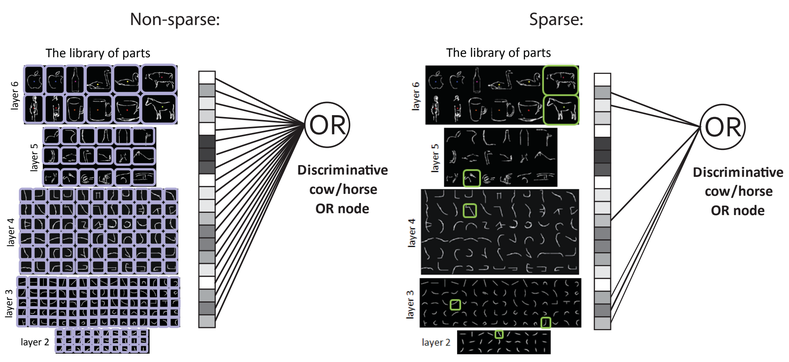
Responses for a horse (sketch) *Responses for a cow (sketch)*

3 Adding discriminative power

- We cast **identification of discriminative parts** and formation of discriminative nodes as a problem of finding a **sparse linear separation** between the cumulative histograms of parts responses.



- Sparsity of vector Θ results in selection of discriminative parts:



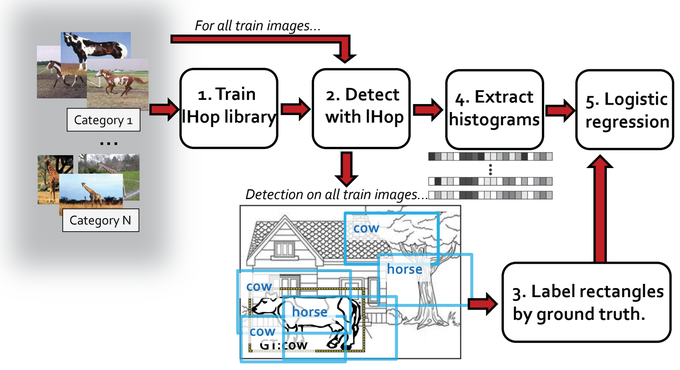
- Learn the posterior over scores $f(\mathbf{h};\Theta)$ by logistic regression:

$$p(\text{horse}|\mathbf{h},\Theta^{(\text{cow,horse})}) = \frac{1}{1+\exp(-f(\mathbf{h};\Theta^{(\text{cow,horse})}))}$$

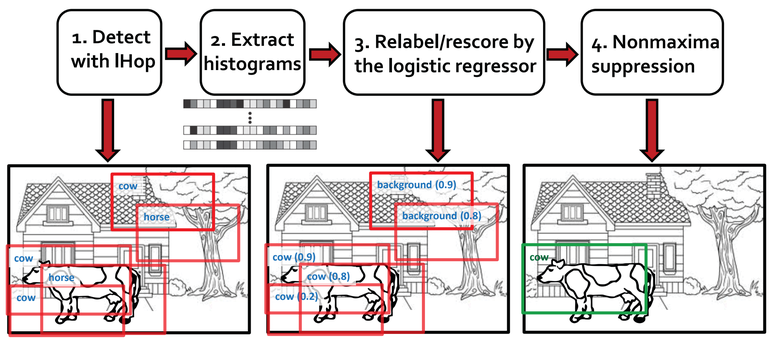
- Sparse Θ is computed via a variational approach from [9].

4. Training / detection

Training stage:



Detection stage:



5. Experimental results

Two experiments designed to evaluate improvement of the proposed discriminative lHop (dlHop) over the baseline lHop [1].

Experiment 1:

- **Goal:** Analyze discrimination between two visually similar categories and the background.
- Dataset composed of Weizman Horses (Borenstein&Ullman, 2008) and Leeds Cows (Leibe et al., 2008):

- Split: 75 images for training, 263 for testing.
- lHop with 7 layers with [6, 33, 161, 180, 93, 104, 2] vocabulary parts.
- PASCAL criterion at overlap >0.3 with ground truth (GT): False positives per experiment and average precision (AP)
- Confusion matrix: Classify by the strongest detection (after nonmax suppression) that overlaps >0.3 by GT.

Confusion matrix

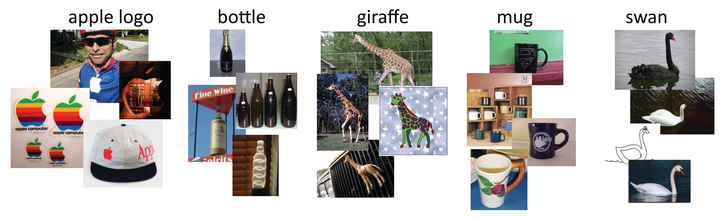
	lHoP (199.8 fp)			dlHoP (166.4 fp)		
	cow	horse	back.	cow	horse	back.
cow	32.44	40.36	27.19	62.57	15.0	22.43
horse	12.25	68.63	19.11	6.50	78.45	15.05
AP	0.31			0.51		

- **dlHop** significantly **reduces confusion** of cows and horses.
- On average, **8% of all parts** were selected for discrimination.

5. Experimental results (cont.)

Experiment 2:

- **Goal:** Demonstrate the improvements in hypotheses rescoring.
- ETHZ shape dataset (Ferrari et al., 2010):

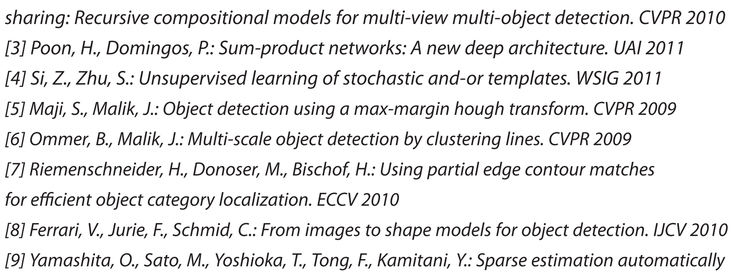


- Half images of category for training and the other half + all images of remaining categories for testing, (5 random splits).
- Evaluation: Detection rate at 1FP per image at 50% overlap
- lHop trained with 7 layers (~525 parts per experiment).

Results of hypothesis rescoring. N_{disc} denotes the number of selected library parts.

	lHoP [1]	dlHoP [N_{disc}] our work	PSM [7]	Hough [8]	w_{ac} [6]	M^2HT [5]	PMK [6]	PMK [7]
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

- dlHop **consistently outperforms** lHop.
- On average, dlHop outperforms, or exhibits performance comparable to, the competing methods.



Example of selected discriminative parts along with the distribution of frequently chosen parts over several repetitions of the experiment.

- Parts were selected from different layers for each category.
- Most parts selected from **layers 3-5**.
- Emphasizing more **global distinctive features** for categorization.

6. Conclusion and future work

- Proposed a simple way to identify **discriminative parts** in a generative hierarchy.
- **Detection improved** by discriminative rescoring.
- The approach **does not hamper** the lHop's **advantages** (sharing, storage, scaling).
- Current approach takes into account **only frequency** of parts.
- **Future work** will explicitly take into account the **positions as well**.

7. References

[1] Fidler, S., Boben, M., Leonardis, A.: Evaluating multi-class learning strategies in a hierarchical framework for object detection. NIPS 2009

[2] Zhu, L.L., Chen, Y., Torralba, A., Freeman, W., Yuille, A.: Part and appearance sharing: Recursive compositional models for multi-view multi-object detection. CVPR 2010

[3] Poon, H., Domingos, P.: Sum-product networks: A new deep architecture. UAI 2011

[4] Si, Z., Zhu, S.: Unsupervised learning of stochastic and-or templates. WSIG 2011

[5] Maji, S., Malik, J.: Object detection using a max-margin hough transform. CVPR 2009

[6] Ommer, B., Malik, J.: Multi-scale object detection by clustering lines. CVPR 2009

[7] Riemenschneider, H., Donoser, M., Bischof, H.: Using partial edge contour matches for efficient object category localization. ECCV 2010

[8] Ferrari, V., Jurie, F., Schmid, C.: From images to shape models for object detection. IJCV 2010

[9] Yamashita, O., Sato, M., Yoshioka, T., Tong, F., Kamitani, Y.: Sparse estimation automatically selects voxels relevant for the decoding of fMRI activity patterns. NeuroImage 2008

This work was supported in part by ARRS Research Programs P2-0214, P2-0095 and ARRS research projects J2-4284, J2-360, J2-2221 and FP7-ICT PACMAN