Object Detection with Grouped Features

Abhilash Srikantha[1,2] and Juergen Gall[1]

[1]Computer Vision Group, University of Bonn
[2]Perceiving Systems Department, Max Planck Institute for Intelligent Systems

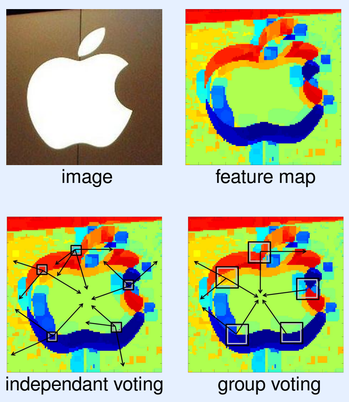
1.Quick Summary

Hough based voting approches model objects by codebooks and their spatial offsets to the center. These codewords are treated independently.

In this work, we propose to model object hypoth--esis on features grouped in local neighborhood.

Conclusions:

- Grouped and individual features are complementary
- Oblique forests for grouped features
- Combining both features yields state-of-the art performance
- Features evaluated on four (RGB,RGBD) datasets



2.Independant Voting

Hough voting scheme: $$p(\mathbf{h}|I) \approx \sum_{\mathbf{y}\in\Omega} p(\mathbf{h}|I(\mathcal{N}(\mathbf{y})))$$

Each patch Pi consists of features I_i, class label c_i and offset to center $\mathbf{d}_i$

$$\{P_i = (I_i, c_i, \mathbf{d}_i)\}$$

Binary tests are based on pixel wise differences of feature I within the patch

$$f_\phi = \begin{cases} 0 & \text{if} \quad I^l(\mathbf{p}) - I^l(\mathbf{q}) < \tau \ , \\ 1 & \text{otherwise} \end{cases}$$

Each node optimizes for maximum gain in classification or regression

$$\Delta G_o = H_o(\mathcal{P}) - \textstyle\sum_{l\in\{0,1\}} \frac{|\mathcal{P}_l|}{|\mathcal{P}|} H_o(\mathcal{P}_l)$$

Each leaf stores class distributions and offsets to object center
A four dimensional parametric space (location x,y; scale s; and aspect ratio a) is spanned for computing object hypotheses

$$p(\mathbf{h}(c,\mathbf{x},s,a)|I(\mathbf{y})) = \frac{1}{|\mathcal{T}|}\sum_{T\in\mathcal{T}} p(\mathbf{h}(\cdot)|L_T(\mathbf{y}))$$

$$p(\mathbf{h}|L_T(\mathbf{y})) = p(\mathbf{d}(\mathbf{y},\mathbf{x},s,a)|c, L_T(\mathbf{y})) \cdot p(c|L_T(\mathbf{y}))$$

3.Group Voting

For this forest, features are leaf assignments based on the above forest
In other words, group features from tree T is histogram of leaves (HOL_T).

$$\{G_i = (\mathrm{HOL}_i, c_i, \mathbf{d}_i)\}$$

Given a leaf L_T of tree T, $HOL_T(L_T)$ is the probability of L_T in HOL_T.
Weights w_T are used to linearly combine between different trees resulting in Oblique forests. Axis aligned forest is a special case where w_T is strictly binary (test depends only on a single tree).

$$f_\phi = \begin{cases} 0 & \text{if} \quad \sum_{T\in\mathcal{T}} w_T \cdot \mathrm{HOL}_T(L_T) < \tau \ , \\ 1 & \text{otherwise} \end{cases}$$

Each leaf stores class distributions and offsets to object center
A four dimensional parametric space (location x,y; scale s; and aspect ratio a) is spanned for computing object hypotheses

$$p(\mathbf{h}(c,\mathbf{x},s,a)|I(\mathcal{N}(\mathbf{y}))) = p(\mathbf{h}(\cdot)|\{L_T(\mathcal{N}(\mathbf{y}))\}_{T\in\mathcal{T}})$$

$$= \frac{1}{|\mathcal{T}_{gr}|}\sum_{T_{gr}\in\mathcal{T}_{gr}} p(\mathbf{h}|L_{T_{gr}}(\{L_T(\mathcal{N}(\mathbf{y}))\}_{T\in\mathcal{T}}))$$

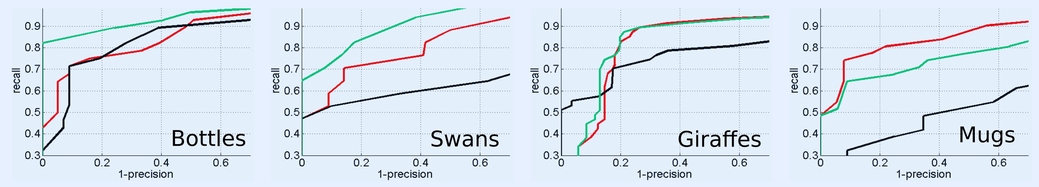
Both grouped and individual features can be linearly combined as

$$p(\mathbf{h}|I(\mathcal{N}(\mathbf{y})),\lambda) \propto\ p(\mathbf{h}|I(\mathcal{N}(\mathbf{y})))^\lambda \cdot p(\mathbf{h}|I(\mathbf{y}))^{1-\lambda}$$

4.Results

support:	7×7			13×13		
depth:	5	10	16	5	10	16
Applelogos	55.0/60.0	**90.0/90.0**	75.0/75.0	15.0/15.0	75.0/75.0	10.0/10.0
Bottles	**92.8/92.8**	89.2/89.2	75.0/75.0	57.1/57.1	71.4/71.4	32.2/32.2
Giraffes	72.3/74.5	78.7/80.8	83.0/83.0	61.7/61.7	**83.0/85.1**	74.5/74.5
Mugs	61.3/61.3	**67.7/74.2**	61.3/61.3	45.2/45.2	61.3/61.3	51.6/51.6
Swans	70.6/76.5	70.6/70.6	58.8/58.8	58.8/58.8	**82.3/88.2**	41.2/58.8

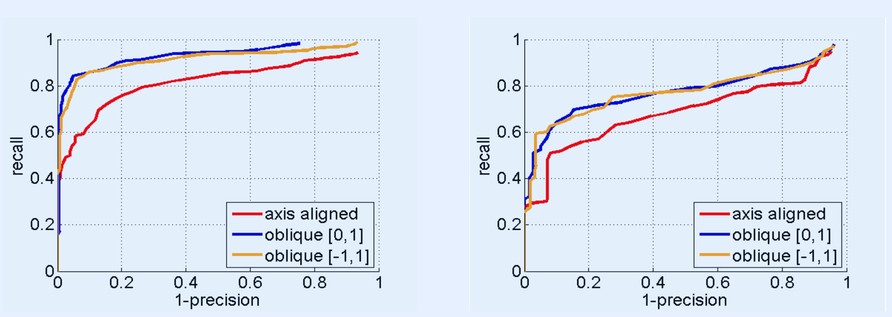
ETHZ Dataset: Setting parameters for the forest of the grouped features (fppi 0.3/0.4)



Precision-Recall plots for independent (red), grouped (black) and best combined (green) features

	Average Precision			optimal λ	Recall at 0.3/0.4 fppi		
	Indi	Group	Comb.		Indi	Group	Comb.
Applelogos	77.8	77.4	**85.4**	0.9	80.0/80.0	90.0/90.0	**90.0/90.0**
Bottles	85.9	84.3	**93.8**	0.7	92.9/96.4	89.2/89.2	**96.4/96.4**
Giraffes	82.6	76.9	**83.4**	0.1	**91.5/93.6**	78.7/80.8	91.5/91.5
Mugs	**84.9**	62.6	84.1	0.1	**90.3/90.3**	67.7/74.2	87.1/90.1
Swans	83.2	63.3	**90.2**	0.6	100/100	70.6/70.6	**100/100**

Performance figures with various settings. Best performing lambda chosen.



Weizmann and INRIA Horse Datasets: Precision-Recall plots for grouped features

	measure	proposed	[1]	[2]	[3]	[4]	[5]	[6]
INRIA	recall	88.0	93.7	92.4	87.3	85.3	×	×
Weizmann	AP	97.2	×	×	×	×	98.0	96.0
Weizmann	recall	94.3	×	×	×	×	95.1	91.5

Recall at 1.0fppi for the combined setting. Parameter set using validation dataset.

Class	RGB	RGBD	Group	bestComb.	λ	Combin.	[7]
bowl	0.231	0.402	0.394	0.423	0.5	0.420	**0.430**
cup	0.123	0.346	0.339	0.358	0.5	**0.357**	0.260
monitor	0.282	0.540	0.530	0.547	0.4	0.547	**0.750**
mouse	0.208	0.282	0.275	0.302	0.4	**0.301**	0.190
phone	0.076	0.163	0.129	0.172	0.3	0.163	**0.180**
keyboard	0.085	0.314	0.283	0.321	0.4	**0.321**	0.170
chair	0.028	**0.208**	0.161	0.211	0.4	0.206	0.140
bottle	0.022	0.178	0.183	0.201	0.2	**0.195**	0.120

VOCB3DO Dataset: Average precision comparison in various settings.

Running times:
- Tree using individual features: 312s, 1.0GB RAM (training) 3.5s (testing)
- Tree using grouped features: 187s, 480MB RAM (training) 13.8s (testing)

5.References

[1] P. Yarlagadda and B. Ommer, "From meaningful contours to discriminative object shape.," in European Conference on Computer Vision, 2012, pp. 776–779.

[2] A. Toshev, B. Taskar, and K. Daniilidis, "Object detection via boundary structure segmentation.," in IEEE Conference on Computer Vision and Pattern Recognition, 2010, pp. 950–957.

[3] P. Yarlagadda, A. Monroy, and B. Ommer, "Voting by grouping dependent parts," in European Conference on Computer Vision, 2010, pp. 197–210.

[4] S. Maji and J. Malik, "Object detection using a max-margin hough transform.," in IEEE Conference on Computer Vision and Pattern Recognition, 2009, pp. 1038–1045.

[5] J. Gall, A. Yao, N. Razavi, L. Van Gool, and V. Lempitsky, "Hough forests for object detection, tracking, and action recognition," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 33, no. 11, pp. 2188–2202, 2011.

[6] J. Shotton, A. Blake, and R. Cipolla, "Efficiently combining contour and texture cues for object recognition.," in British Machine Vision Conference, 2008.

[7] A. Janoch, S. Karayev, Y. Jia, J.. Barron, M. Fritz, K. Saenko, and T. Darrell, "A category-level 3d object dataset: Putting the kinect to work," in Consumer Depth Cameras for Computer Vision, pp. 141–165. Springer, 2013.

asrikantha@informatik.iai.uni-bonn.de
gall@informatik.iai.uni-bonn.de