Image Classification Using Object Detectors

Thibaut Durand[1], Nicolas Thome[1], Matthieu Cord[1], Sandra Avila[1,2]
(1) Université Pierre et Marie Curie, UPMC-Sorbonne Universities, LIP6, 4 place Jussieu, 75005, Paris, France
(2) Federal University of Minas Gerais, NPDI Lab – DCC/UFMG, Belo Horizonte, MG, Brazil

Contributions

▷ New method for spatially pooling response maps for object detectors to create a discriminative and compact image signature

▷ Combine our representation with BoW-like representations

New spatial pooling strategy

▷ 2 object detectors :

- Latent SVM object detectors [3] for most of the blobby objects
- Texture classifier by Hoiem [4] for more texture- and material-based objects/regions

▷ **Final image representation** Z concatenates the aggregation operator, denoted as $aggr(r, c)$ for each detector c and regions r :

$$Z = [aggr(r,c)]_{(r,c)\in\{1;N_r\}\times\{1;N_c\}}$$

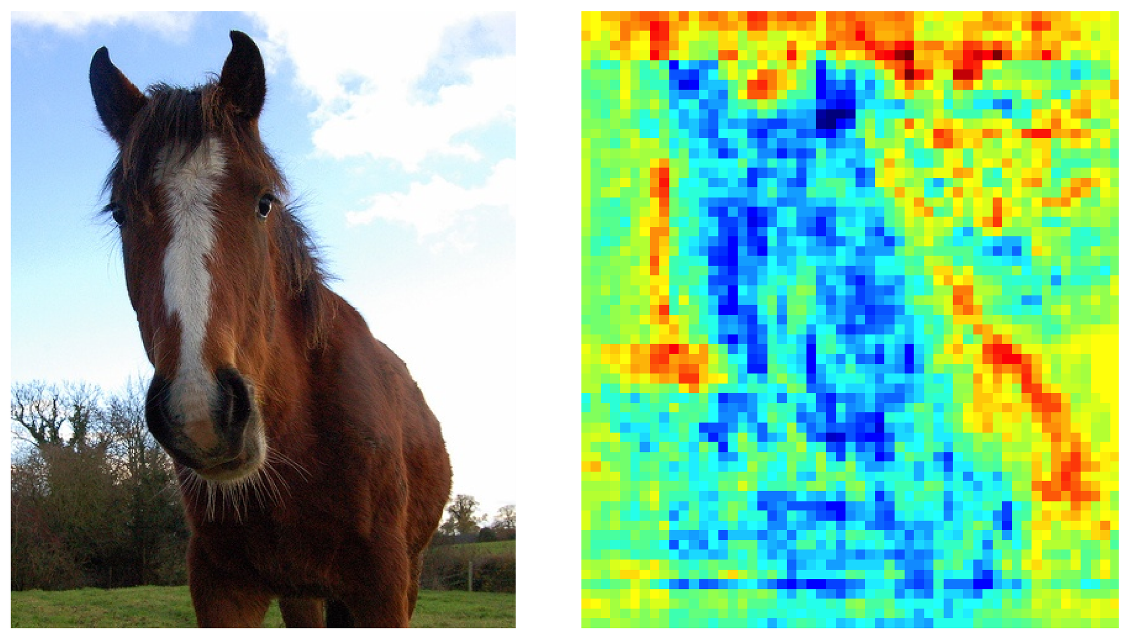
N_c : number of detectors
N_r : number of spatial regions

$$aggr(r,c) = \begin{cases} sum(r,c) & \text{if } c \text{ is a texture} \\ n\text{-}max(r,c) & \text{otherwise} \end{cases}$$

Dimension : $N_r \times (n \times N_{obj} + N_{text})$
N_{obj}: number of object detectors
N_{text}: number of texture detectors

▷ **Spatial pooling with n maximums**
Extract a vector of size n by taking the n bounding boxes with the n largest detection scores
→ keeping information about the number of objects of class c present in region r

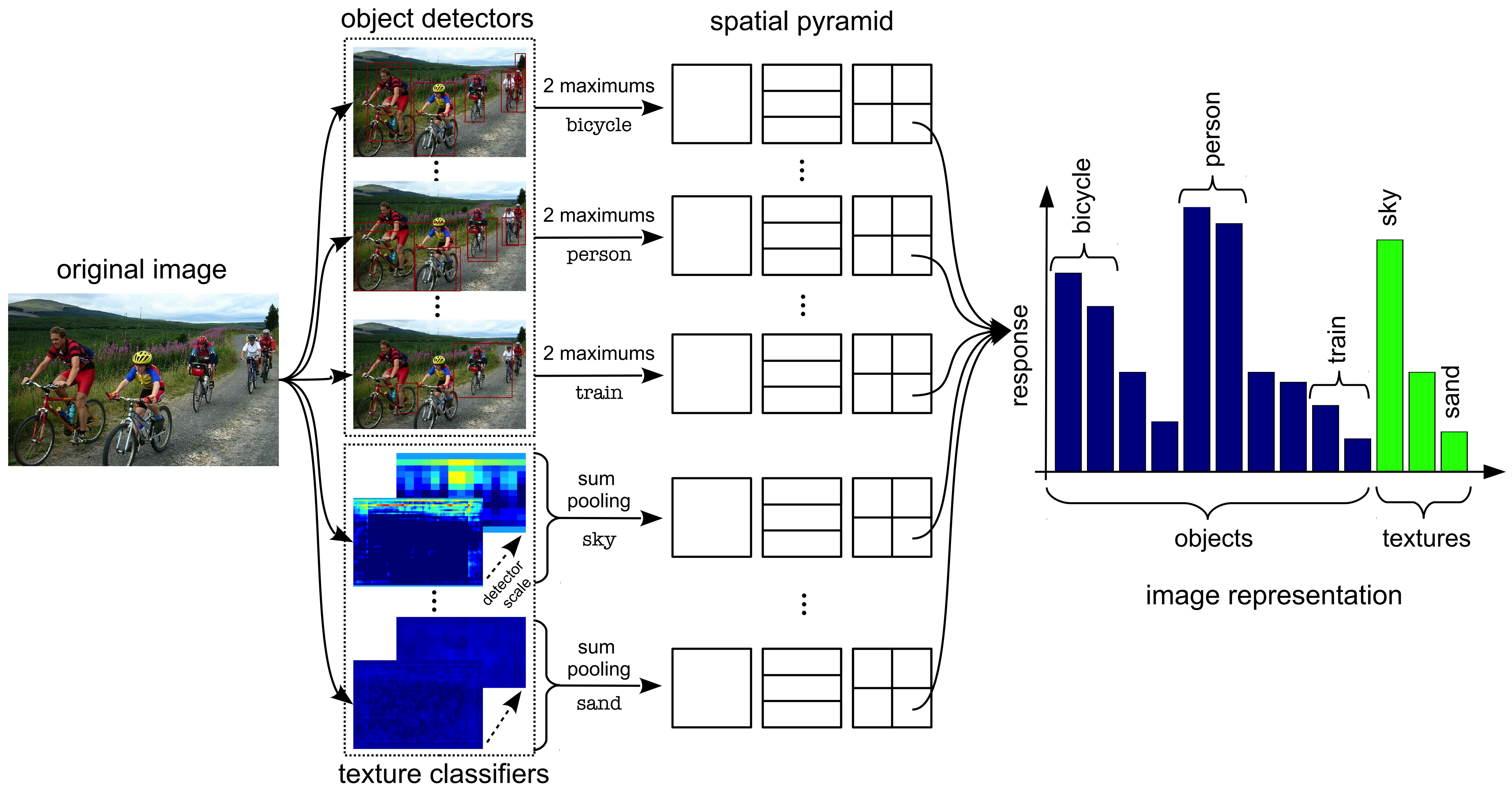
▷ **Different pooling for objects and textures**



Original image (left) and response map (one for sky classifier (right - high values in red)

→ use sum-pooling for texture classifiers

Proposed pipeline



Results

◇ Dataset : PASCAL VOC 2007 (20 classes)
◇ Spatial Pyramid Matching : $1\times 1, 2\times 2, 3\times 1$
◇ Classification performance : Mean Average Precision (MAP)
◇ BN : 4,096 visual words, 2 bins, $\lambda_{min} = 0.4$ and $\lambda_{min} = 2.0$, $s = 10^{-3}$
◇ Fisher Vector : 256 Gaussians
◇ Late Fusion : $\alpha = 0.5$

◇ "One-versus-all" SVM classifier with RBF kernel
◇ Object detectors :

- 20 latent SVM object detectors [3] which correspond to 20 object categories of the VOC 2007
- 6 texture classifiers of Object Bank [1]: *rock-stone, sand, water, sky, grass* and *building-edifice*

	MAP	plane	bicycle	bird	boat	bottle	bus	car	cat	chair	cow
OURS	59.0	64.7	75.6	32.1	62.7	48.2	70.3	83.8	49.3	57.2	48.4
BNFV	60.3	79.5	65.6	53.6	72.1	32.7	66.0	79.0	59.7	54.5	43.0
LF	**67.6**	80.8	78.8	55.4	73.8	52.2	76.6	86.4	64.1	62.1	55.2
		table	dog	horse	moto	person	plant	sheep	sofa	train	tv
OURS		52.4	34.2	76.9	68.1	87.7	36.6	44.4	51.8	71.3	63.8
BNFV		60.0	46.8	78.6	64.8	84.5	31.2	45.3	54.6	78.5	55.1
LF		66.6	49.7	83.4	74.7	89.8	37.9	50.7	64.1	80.9	68.9

Image classification MAP (%) on VOC 2007 dataset (LF: Late Fusion, OURS: our signature)

Method	BNFV	[5]	[6]	[7]	OURS	LF
MAP (%)	60.3	61.7	66.3	66.6	59.0	**67.6**

State-of-the-art results (MAP %) on PASCAL VOC 2007 dataset

Combination

▷ Explore the combination between our signature and low-level representations : BossaNova (BN) and Fisher Vectors (FV) [2]

▷ Combination by late fusion – learn individually each classifiers and compute a linear combination :

$$f(x) = \alpha f_{ours}(x) + (1-\alpha) f_{BNFV}(x) \qquad (1)$$

f_{ours} : classification score for ours signature
f_{BNFV} : classification score for BossaNovaFisher

Conclusion

▷ For texture classifier, sum-pooling is more appropriate than max-pooling

▷ Good results with a compact image representation

▷ The combination with low-level representations outperforms state-of-the-art performances

Acknowledgments

IEEE Signal Processing Society ®

References

[1] L.-J. Li, H. Su, E. Xing, and L. Fei-Fei, "Object bank: A high-level image representation for scene classification & semantic feature sparsification," in *NIPS*, 2010.

[2] S. Avila, N. Thome, M. Cord, E. Valle, and A. de A. Araújo, "Pooling in image representation: The visual codeword point of view," *Computer Vision and Image Understanding*, 2013.

[3] P. Felzenszwalb, R. Girshick, D. McAllester, and D. Ramanan, "Object detection with discriminatively trained part-based models," in *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2010.

[4] D. Hoiem, A. Efros, and M. Hebert, "Automatic photo pop-up," *ACM Transactions on Graphics*, 2005.

[5] K. Chatfield, V. Lempitsky, A. Vedaldi, and A. Zisserman, "The devil is in the details: an evaluation of recent feature encoding methods," in *BMVC*, 2011.

[6] J. Sánchez, F. Perronnin, and T. E. de Campos, "Modeling the spatial layout of images beyond spatial pyramids," *Pattern Recognition Letters*, 2012.

[7] Y. Su and F. Jurie, "Improving image classification using semantic attributes," *International Journal of Computer Vision*, 2012.